arXiv:2305.02389v2 [stat.ME] 3 Jun 2023

快速广义函数主成分分析

Andrew Leroux

生物统计系　　统计学与信息学，科罗拉多大学安舒茨医学院　　us

Ciprian M. Crainiceanu 生物统计系，约翰斯·霍普金斯彭博公共卫生学院

Julia Wrobel

科罗拉多大学安舒茨医学院生物统计与信息学系

2023年6月6日

摘要

我们提出了一种新的快速广义函数主成分分析（fast-GFPCA）算法，用于非高斯函数型数据的降维。该方法包括：（1）在函数域内对数据进行分箱；（2）在每个分箱中拟合局部随机截距广义线性混合模型，以获得个体特异性函数线性预测值的初始估计值；（3）使用快速函数主成分分析对线性预测值进行平滑，并获得其特征函数；（4）以线性预测值的特征函数为条件，估计全局模型。大规模模拟研究表明，fast-GFPCA 的性能不逊于或优于现有最先进的方法；其速度比现有通用广义函数主成分分析方法快几个数量级；并且能够随着观测曲线数量和每条曲线的观测数增加而良好扩展。该方法的提出和应用源于对 NHANES 2011—2014年研究中利用可穿戴加速度计获得的活动/非活动身体活动模式的研究。任何熟悉混合模型软件的用户都可以实现该方法，不过为方便起见，我们提供了 R 包 fastGFPCA。

关键词：函数型数据，函数主成分分析，广义函数主成分分析

1 引言

函数型数据分析（FDA）Ramsay 和 Silverman（2005年）为时间序列或图像等具有复杂结构的数据提供了丰富的分析技术。这类数据通常具有高维特征（每个函数有许多重复观测），并呈现复杂的非平稳变化模式。函数型数据的函数主成分分析（FPCA）Jones 和 Rice（1992年）；Rice 和 Silverman（1991年）；Staniswalis 和 Lee（1998年），是主成分分析（PCA）Hotelling（1933年）；Jolliffe（1982年）；Pearson（1901年）在函数型数据上的对应方法，也是分析此类数据时首选的降维技术。FPCA 与 PCA 的主要区别在于，函数型数据：（1）可能存在显著的测量误差；（2）在定义域的每个点上都以相同的计量单位表示；（3）其函数域具有顺序性和自然距离（例如时间顺序和距离）。

关于函数主成分分析（FPCA）的研究文献十分丰富，涵盖估计和推断两方面。总体而言，针对高斯数据的函数主成分分析方法包括对协方差函数进行平滑 Staniswalis 和 Lee (1998年)；Xiao 等人 (2016)；Yao 等人 (2003)，或在明确的似然假设下采用基于模型的方法进行估计 van der Linde (2008年)；James 等人 (2000年)。针对稀疏或非规则观测数据 J. 和 Paul (2009年)；Yao 等人 (2005年)；Xiao 等人 (2018)、多变量函数数据 Happ 和 Greven (2015年)；Chiou 等人 (2014)；Li 和 Xiao (2021年)，以及多层函数数据 Di 等人 (2009年)；Cui 等人 (2022年) 的扩展方法也已问世。尽管函数主成分分析领域的文献相当丰富，但高质量的开源软件实现却为数不多。Xiao 等人 (2021, 2016) 提出的协方差平滑 FACE 方法，由 Goldsmith 等人 (2020) 在 `refund` 包中实现，是目前针对规则观测数据估计函数主成分分析最快的方法。对于大型数据，基于似然的方法所需的计算时间要长得多。

这里我们关注将函数主成分分析方法扩展到非高斯结局（例如二元或计数数据），并将其称为广义函数主成分分析（GFPCA）。更确切地说，我们关注

的是沿主要变异方向分解个体特异性潜在函数线性预测变量变异的方法。与函数主成分分析相关的大量已发表论文相比，广义函数主成分分析的论文要少得多。少数例外包括单层方法的 Hall等（2008）；van der Linde（2009）；Gertheiss 等人（2017）；Weishampel 等人（2023），以及多层广义函数主成分分析的 Chen等（2013）；Goldsmith等（2015）。遗憾的是，与函数主成分分析相比，广义函数主成分分析的软件实现问题更为突出。事实上，大多数已发表的方法要么没有配套软件，要么在处理大型数据时速度极慢。此外，现有方法要求预先指定用于估计主成分的主成分数量和基函数数量。在将广义函数主成分分析应用于新数据集时，评估这些关键输入参数的敏感性至关重要。

随着规模和复杂度不断增加，收集非高斯函数型数据的研究数量日益增多，因此需要快速且可扩展的方法。例如，考虑在大型流行病学研究（如美国国家健康与营养调查（NHANES）和 UK Biobank）中，通过可穿戴加速度计获取的分钟级别身体活动数据（Doherty et al.，2017）。在许多应用中，人们关注的是一天中每分钟处于活动状态（编码为 1）或非活动状态（编码为 0）的模式。因此，在研究参与者层面，数据是一个在 1440 分钟（一天的分钟数）上观测到的函数，其在每个时间点的取值均为活动或非活动。NHANES 包含数万名研究参与者的此类数据，而 UK Biobank 则包含近 100,000 名研究参与者的数据。我们的目标是提供一种可扩展的广义函数主成分分析方法，通过提取这些函数在线性预测空间中的正交变异方向，帮助解释和分析 NHANES 加速度测量研究中的大规模非高斯函数型数据。

最近，Weishampel 等人（2023）提出了一种方法，可用于估计一般指数族结局的广义函数主成分分析。该方法速度很快，可视为一种替代方案，但在关键方面与 fast-GFPCA 不同。重要的是，在我们的

数据应用方面，Weishampel 等人（2023）的方法可能导致估计的潜在函数出现无限偏差。我们将在第 2.5 节进一步讨论这一点，但由于该方法不适用于我们的数据应用，因此不对其进行重点讨论。除 Weishampel 等人（2023）之外，能够容纳多种类型指数族数据的通用广义函数主成分分析软件运行缓慢。然而，特定类型的结局也有令人鼓舞的进展。例如，Wrobel 等人（2019）开发了一种非常快速且高效的二元函数主成分分析过程，该过程使用期望最大化算法来优化二项函数主成分分析似然的变分近似。该论文随附`registr` 软件包 Wrobel（2018）；Wrobel 和 Bauer（2021）。据我们所知，这是唯一公开可用、对于大型数据集而言速度快且可行的广义函数主成分分析实现，因此我们将 fast-GFPCA 方法与 `registr::bfpca()` 函数进行比较。此外，在数据集较小的模拟情景中，我们将 fast-GFPCA 与 Gertheiss 等人（2017）所描述的两步广义函数主成分分析实现进行比较；该实现以 `registr::gfpca twoStep()` 函数的形式包含在 `registr` 软件包中。这种两步方法可用于建模多种指数族结局，但对于大型数据集而言速度慢到无法接受。当前工作通过提供一种广义函数主成分分析通用方法，为相关文献作出了实质性贡献；该方法可轻松推广至我们所关注的函数型回归模型，而目前针对大规模数据尚无快速实现。

-

具体而言，我们提出了一种新方法 fast-GFPCA，其理念和实现策略与大多数其他方法截然不同。fast-GFPCA 的优点包括：（1）适用于任何类型的非高斯结局，而不仅仅是二元结局；（2）速度比大多数其他通用型 GFPCA 方法快几个数量级；（3）能够轻松处理缺失数据；（4）可以轻松扩展以纳入协变量；（5）可以推广至多层、纵向或结构化函数型数据。我们将简要讨论这些扩展，但具体细节留待未来工作。

本文其余部分安排如下。第 2 节介绍快速GFPCA（fast-GFPCA）方法。接下来，在第 3 节中，我们将 fast-GFPCA 应用于

使用美国国家健康与营养调查（NHANES）2011—2014年各轮数据，通过可穿戴加速度计获得的活动/不活动剖面。随后，我们在第4节的模拟研究中展示 fast-GFPCA 方法的实用性。最后，我们在第 5 节进行讨论。

2 方法

观测数据结构的类型为 $\{s_j, Z_i(s_j)\}$，其中 $Z_i(s_j)$ 是受试者 $i \in 1$ 在点 $s_j \in S$ 处针对 $j \in 1$ 的非高斯函数观测，$\ldots, N$， J。我们假设这些 $\{s_j, Z_i(s_j)\}$ 对是来自连续过程 $\{Z_i(s) : s \in S\}$ 的离散实现，满足：(1) $g[E\{Z_i(s)\}] = \beta(s) + b_i(s)$，其中 $g(\cdot)$ 是适当的链接函数， $\beta(s)$ 是线性预测空间中的总体均值函数， $b_i(s)$ 是线性预测空间中个体相对于总体均值的偏差；(2) $b_i(s) \sim \text{GP}(0, K_b)$ 是一个均值为零、协方差算子为 K_b 的高斯过程。我们的目标是沿潜在过程 $b_i(s)$ 的主要变异方向分解其变异（获得 FPCA 分解），并在给定这些变异方向的条件下估计 $b_i(s)$。尽管 $b_i(s)$无法直接观测，我们可以利用 Karhunen-Lo`eve (KL) 展开 $b_i(s) = \sum_{k=1}^{\infty} \xi_{ik}\phi_k(s)$，其中 $\phi_k : S \to \mathbb{R}$ 是正交归一特征函数，满足 $\int_S \phi_k^2(s)ds = \lambda_k$， $\lambda_1 \geq \lambda_2 \geq \ldots$ 是特征值，而 $\xi_{ik} \sim N(0, \lambda_k)$ 是研究参与者 i 中相互独立的受试者特异性得分，且特征值为 k。由此得到广义函数主成分分析模型,

$$g\left(E[Z_i(s)]\right) = \eta_i(s) = \beta_0(s) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(s)\ . \tag{1}$$

这与经典的函数主成分分析（FPCA）模型非常相似，唯一的区别在于噪声不一定服从高斯分布。

2.1 快速GFPCA

对于观测数据为非高斯分布的情况，我们提出以下快速广义函数主成分分析算法（fast-GFPCA），以对潜在过程进行函数主成分分析（FPCA）：(1) 沿观测函数域 $\mathcal{S} = 1,$ ……， J将数据分为 L个可能存在重叠的分箱；(2) 在每个分箱中分别估计带随机截距的局部广义线性混合模型，以获得线性预测尺度上的受试者特异性估计值；(3) 对步骤 (2) 获得的受试者特异性估计值进行函数主成分分析；(4) 使用步骤 (3) 获得的估计特征函数，在原始数据的分辨率下重新估计广义函数主成分分析。上述每个步骤都很简单，不过为方便起见，我们提供了 `R` 软件包 `fastGFPCA`，其介绍见第2.3节。第 2.2 节将详细介绍每个步骤。

2.2 估计算法

现在，我们进一步详细介绍快速广义函数主成分分析算法的四个步骤；算法描述见补充材料的 S-1 节。

步骤1： 选择如何对数据进行分箱。分箱数量（L）和分箱宽度（w_l）的选择将依据可识别性以及对潜在过程复杂度的假设来确定。具体而言，假设我们选择 L 个分箱，其中 m_l 是第 $l = 1$、……、L 个分箱的中点。除定义域 S 的边界处外，我们使用对称分箱。对于定义域上按规则间隔网格观测的数据，第 l^{th} 个分箱包含定义域取值为 $\mathcal{S}_l = \{s_{m_l - w_l/2}, \ldots, s_{m_l}, \ldots, s_{m_l + w_l/2}\}$ 处的数据，共有 $w_l + 1$ 个点。随后将数据分箱为 $[\{Z_i（s_j） \quad l\}、\ 1 \leq i \leq N、\ j \in \mathcal{S}_l、\ 1 \leq l \leq L]$。

如果数据具有周期性（如我们的数据应用），分箱可能会跨越边界。例如，对于分钟级别的加速度计数据，如果令 $m_1 = 1$（活动时间 00:00-00:01）和 $w_l = 6$，那么 $\mathcal{S}_1 = \{1438, 1439, 1440, 1, 2, 3, 4\}$（活动时间 23:57-00:05）。当数据不具有周期性时，我们建议构造分箱，使其为 $\mathcal{S}_l = \{\min(s_{m_l - w_l/2}, s_1) \ldots, s_{m_l}, \ldots, \max(s_{m_l + w_l/2}, s_J)\}$，这样每个分箱中的数据点最少可达 $w_l/2 + 1$ 个。

步骤2： 在每个分箱中拟合一个局部 GLMM。具体而言，在每个分箱 $l = 1, \ldots, L$ 中，我们估计以下形式的独立模型： $g[E\{Z_i(s_j)\}] = \beta_0(s_{m_l}) + b_i(s_{m_l}) = \eta_i(s_{m_l})$，其中 $s_j \in \mathcal{S}_l$。这里，$\beta_0(s_{m_l})$ 是固定效应均值， $b_i(s_{m_l})$ 是在分箱中心 $\widehat{}$ 处取值的随机截距， s_{m_l}。通过这些模型，我们得到总体均值 $\beta_0(s_{m_l})$ 的估计值，以及 $\widehat{}$ 个体特异性随机效应 $b_i(s_{m_l})$ 的预测值。这些预测值并非 位于函数被观测到的原始网格上，而是位于中点 $\{s_{m_1}, \ldots, s_{m_L}\} \subset \mathcal{S}$。重要的是，该模型存在设定错误，因为它假定每个分箱内的效应为常数，而实际上， $\beta_0(\cdot)$ 和 $b_i(\cdot)$ 都会随 $\mathcal{S}$ 平滑变化。这可能导致 $\beta_0(s_{m_l})$、 $b_i(s_{m_l})$ 的有偏估计，从而可能使我们对 K_b 的估计量产生偏差。在分箱宽度选择合理的情况下，这种偏差大多会被估计协方差算子的特征值吸收，而特征函数仍能得到良好估计。由于该方法的关键是在步骤 3 中准确估计特征函数，而不一定要准确估计特征值，因此这并不会影响该方法。下文将暂时偏离快速广义函数主成分分析算法的描述，更详细地讨论偏差问题，因为我们认为，这样的讨论有助于深入理解所提出的方法为何在实践中有效且合理。

分箱引起的偏差对 K 估计值的影响讨论b 我们认为，分箱过程会导致在每个分箱中点处评估的潜在函数估计值产生偏差，进而导致协方差函数估计值产生偏差 K_b。当假设观测数据沿定义域服从连续均匀分布时，最容易说明这一点。为便于说明，假设定义域为 $S = [0, 1]$。由此可知，点的分布密度为 $f_s(s) = 1$, $s \in [0, 1]$。此外，给定 $s \in S_l$ (s 位于第 l^{th} 个分箱中)，点服从独立同分布的均匀分布，其条件密度为 $f_{s|l}(s) = |S_l|^{-1}$ （区间宽度的倒数）。

上标 $_{\text{bin}}$ 表示错误设定模型下的目标估计量，以区别于“真实”的潜在过程。步骤2中的错误设定模型估计

$\eta_i^{\text{bin}}(s_{m_l}) = E[\beta_0(s) + b_i(s)|\{\xi_{ik}: \ 1 \le k \le K\}, \ s \in \mathcal{S}_l] = E[\beta_0(s)|s \in \mathcal{S}_l] + E[b_i(s)|\{\xi_{ik}: 1 \le k \le K\}, \ s \in \mathcal{S}_l]$，其偏差可分为总体偏差 $\left[\beta_0^{\text{bin}}(s_{m_l})\right] = \beta_0(s_{m_l}) - E[\beta_0(s)|s \in \mathcal{S}_l]$ 和受试者特异性偏差 $\left[b_i^{\text{bin}}(s_{m_l})\right] = b_i(s_{m_l}) - E[b_i(s)|\xi_{ik}, \ s \in \mathcal{S}_l]$两个部分。

fast-GFPCA 的关键在于能够在步骤 3 中获得协方差算子 $K_b = \text{Var}[b_i(s)]$ 的合理估计值。由于加性偏差不会影响步骤 3 中协方差算子的估计量，因为数据为（即 $\text{Cov}(a + X, Y) = \text{Cov}(X, Y)$），因此我们重点关注受试者特异性偏差的影响，其表示为

$$\begin{aligned}
\text{Bias}\left[b_i^{\text{bin}}(s_{m_l})\right] &= b_i(s_{m_l}) - E[b_i(s)|\xi_{ik}, s \in \mathcal{S}_l] \\
&= \sum_{k=1}^{K} \phi_k(s_{m_l})\xi_{ik} - E[\sum_{k=1}^{K} \phi_k(s)\xi_{ik}|\xi_{ik}, s \in \mathcal{S}_l] \\
&= \sum_{k=1}^{K} \phi_k(s_{m_l})\xi_{ik} - \sum_{k=1}^{K} \xi_{ik} \int_{s \in S_l} \phi_k(s) f_{s|l}(s) ds \\
&= \sum_{k=1}^{K} \phi_k(s_{m_l})\xi_{ik} - \sum_{k=1}^{K} \xi_{ik} |S_l|^{-1} \int_{s \in \mathcal{S}_l} \phi_k(s) ds
\end{aligned}$$

其中 $|S_l|^{-1}$ 作为归一化常数，使得 $\lim_{w_l \to 0} |S_l|^{-1} \int_{s \in \mathcal{S}_l} \phi_k(s) ds = \phi_k(s_{m_l})$。也就是说，随着分箱宽度趋于 0，我们恢复出在当前分箱中点处取值的特征函数。然后

$$\begin{aligned}
\text{Cov}\left[b_i^{\text{bin}}(s_{m_u}), b_i^{\text{bin}}(s_{m_v})\right] &= \sum_{k=1}^{K} \lambda_k \left[|S_u|^{-1} \int_{\mathcal{S}_u} \phi_k(s) ds\right] \left[|S_v|^{-1} \int_{\mathcal{S}_v} \phi_k(s) ds\right] \\
&\approx \sum_{k=1}^{K} \{\lambda_k |S_u|^{-1} |S_v|^{-1}\} \phi_k(s_{m_u}) \phi_k(s_{m_v}) \ .
\end{aligned}$$

最后一个近似与中值定理有关。该定理指出，只要 $|S_u|$、$|S_v|$ 相对于 ϕ_k 中的曲率而言不大，$\frac{1}{|S_u|} \int_{S_u} \phi_k(s) ds \approx \phi_k(s_{m_u})$，且 $\frac{1}{|S_v|} \int_{S_v} \phi_k(s) ds \approx \phi_k(s_{m_v})$。如果函数 $\phi_k(\cdot)$ 在这些区间中的任一区间上为常数或线性函数，则该近似实际上就是等式。该方程表明，通过对 $\text{Cov}\left[b_i^{\text{bin}}(s_{m_u}) \ b_i^{\text{bin}}(s_{m_v})\right]$ 进行对角化，我们可以得到特征向量 $\phi_k(\cdot)$ 的较好近似，但无法得到

特征值 λ_k 的较好近似。事实上，特征值会按 $|S_u||S_v|$ 重新缩放；在使用固定组距时，$|S_u||S_v|$ 是近似区间长度的平方。这也解释了为什么协方差估计中的所有偏差都会被特征值吸收，并隐含地传递给得分。因此，如果我们将这一步得到的线性预测值进行对角化，就会得到无偏的特征函数，但协方差、特征向量、得分和个体特异性轨迹都会存在偏差。

步骤 3： 使用线性预测尺度上的预测响应 $[\{\widehat{\eta}_i(s_j)\ l\},\ 1 \le i \le N,\ j \in \mathcal{S}_l]$ 通过 `refund::fpca.face()` 函数中实现的快速协方差估计（FACE）方法，估计 $K_b(u,\ v) = \text{Cov}\{\widehat{\eta}_i(u)\ \widehat{\eta}_i(v)\}$（针对 $u,\ v \in \{s_{w_1}, \ldots, s_{w_L}\}$）。$\widehat{\ }$ 获取协方差算子$\widehat{\ }$的估计特征函数 $[\{\phi_k(s_{w_l})\},\ 1 \le l \le L,\ 1 \le k \le K]$，其中特征函数的数量 K通过例如方差解释百分比（如 95%、99%）来选定。

步骤 4： 在步骤 3 中获得的基函数条件下估计广义函数主成分分析模型。由于基函数是在与观测数据不同的网格上估计的，我们首先将每个特征函数投影到步骤 3 算法的 FACE 组件中使用的丰富 B 样条基上。这样便可估计数据最初采样位置处的特征函数。完成这些投影后，fast-GFPCA 可表示为以下广义线性混合模型（GLMM）

$$g(E[Z_i(s)]|\{\hat{\phi}_k(s) : 1 \le s \le J, 1 \le k \le K\}) = \beta_0(s) + \sum_{k=1}^{K} \xi_{ik}\hat{\phi}_k(s)\ , \tag{2}$$

其中，$\xi_{ik} \sim N(0,\ \sigma_k^2)$ 相互独立，$\beta_0(s)$ 是一个未指定的平滑函数。也就是说，给定 $\phi_k(\cdot)$ 的估计值，这是一个具有 K 个不相关随机斜率的广义线性混合模型。使用具有不相关随机斜率的主成分分解，可以简化随机效应协方差结构，使得只需估计 K 个方差参数，从而提高计算效率。

如果我们假设 $\beta_0(s)$ 具有参数形式，则可以使用任何广义线性混合模型软件。对于以非参数方式建模 $\beta_0(s)$ 的情况，我们使用 Wood 等人（2017）的 `mgcv::bam()`函数，并结合快速 REML 平滑参数选择以及 Li 和 Wood（2020）中的参数 `discrete=TRUE`，来估计模型 2。该方法具有很高的计算效率；示例代码见第2.3节。也有用于估计加性广义线性混合模型的其他软件，在某些情况下可能具有更快的运行速度和更高的内存效率；具体而言，Wood（2011）的 `mgcv::gamm()` 函数和 Wood 等人（2013）的 `gamm4::gamm4()` 函数分别为 `mgcv`与 `nlme` Pinheiro 和 Bates（2000）软件包以及 `lme4` Bates 等人（2015）软件包提供了接口。

2.3 快速GFPCA：示例代码

快速广义函数主成分分析算法的一大优点是，任何熟悉混合模型软件的人都可以实现它。以下代码展示了如何将快速广义函数主成分分析算法应用于第 3 节所述应用中使用的二元NHANES活动/非活动剖面数据的一个子集。作为示例，我们使用重叠窗口，分箱宽度为 $w_l = 10$，适用于 $l = 1$、.....、1440（ w_l和 l均以分钟为单位）。代码的总体组织遵循快速广义函数主成分分析算法的四个步骤。使用配套的 `fastGFPCA` 软件包，也可以通过一行代码运行此算法。

library("refund");library("tidyverse"); library("lme4") df <-read.rds("
NHANES_example.rds") # 读取数据 J <-length(unique(df$index)) # 维度

步骤1：分箱决策 bin_len <-10 # 分箱宽度 (w) s_m <-1:J # 分箱中点 s_
{m_l}

步骤2：拟合局部分箱广义线性混合模型 fit_ls <- vector(mode="list", length=J)
空列表，用于存储结果 for(j in s_m){ # 遍历分箱 # 获取与当前分箱关联的索引 S_l #
注意：此数据是循环的，因此我们会跨越定义域查看 # 使用取模（%%）函数

```
sind_j <- (j-bin_len_w/2):(j+bin_len_w/2) %% 1440 sind_j[sind_j == 0] <- 1440 # 子集筛选当前索引 df_j <- df %>% 筛选(index %in% sind[sind_j]) # 拟合局部分箱广义线性混合模型： # g(E[Z_i(s)]|s_j \in S_l) = \beta_0(s_{m_l}) + b_i(s_{m_l}) fit_j <- glmer(value ~ 1 + (1|id), data=df_j, family=二项式) # 存储结果（id、\hat{eta}_i(s_{m_l})、s_{m_l}） fit_ls[[j]] <-data.frame("id" = 1:N, "eta_i" = coef(fit_j)$id[[1]],"s_m" =j)
```

```
} ## 按行绑定列表元素 fit_df <- bind_rows(fit_ls)
```

```
## 步骤 3：对分箱后的潜在估计值进行函数主成分分析 fpca_latent <- fpca.face(matrix(fit_df$eta_i, N, J, byrow=FALSE), pve=0.95, argvals=sind, knots=20,lower=0)
```

```
## 步骤 4：在完整数据上重新拟合广义线性混合模型
## 注意，这里不需要对特征函数进行插值，因为分箱中点包含了原始观测点 # 特征函数数据框（取前四个） phi_mat <- data.frame("index" = s_m, fpca_latent$efunctions[,1:4]) colnames(phi_mat)[2:5] <- paste0("Phi", 1:4) # 合并数据 df <- df %>% left_join(phi_mat, by="index") %>% mutate(id_fac=factor(id)) # 使用 mgcv::bam() 重新拟合模型 fit <- bam(value ~s(index,bs="cc",k=20) +s(id_fac,by=Phi1,bs="re") + s(id_fac,by=Phi2,bs="re") +s(id_fac,by=Phi3,bs="re") + s(id_fac,by=Phi4,bs="re"), data=df, family=binomial, discrete=TRUE, method="fREML")
```

为便于展示，代码假定数据先按受试者排序，再按函数定义域排序。如果数据未按此方式排序，则需要对步骤 3 和步骤 4 稍作修改，或者在运行代码前重新排列数据。此外，要拟合正确的模型，受试者标识符必须是因子变量，这也是步骤 4 中创建变量 `id fac` 的原因。不熟悉 `mgcv` 软件包的用户可能会对步骤 4 中调用 `mgcv::bam()` 时所使用的语法感到困惑。表达式 `s(id fac, by=Phi1,bs="re")` 构造了相互独立且服从正态分布的随机斜率。同一个随机

lme4 中的效应设定将是

```
fit <- glmer(value~1 + (0+Phi1|id) + (0+Phi2|id) + (0+Phi3|id)+(0+Phi4|id),
data=df、 family=binomial)、
```

然而， lme4::glmer()无法对 $\beta_0(t)$进行非参数建模，且必须为$\beta_0(s)$指定参数形式。我们希望，看到完整代码能够：(1) 展示所提出的方法实现起来有多简单；(2) 提供一个模块化且可修改的平台，以便用于需要特定调整的类似情形；(3) 支持这样一种理念：没有配套软件，分析方法就算不上真正的方法。

为方便用户，我们开发了 fastGFPCA，这是一个 R 软件包，可在 GitHub 上通过 https://github.com/julia-wrobel/fastGFPCA获取，其中封装了实现 fast-GFPCA 的代码。软件包功能的详细说明，包括二项式和泊松分布函数型数据的示例，收录于该软件包关联的 fastGFPCA指南中。简而言之，软件包的主要函数是fast gfpca()。函数参数 overlap 和 binwidth 允许用户为数据选择分箱宽度，并决定是否构建重叠窗口。参数 family 会传递给函数 lme4::glmer() 和 mgcv::bam() ，用于指定指数族分布和链接函数，例如使用语法family = binomial(link = "logit")。 fast gfpca()函数会返回一个 "fpca" 类对象，结果可借助 Wrobel et al. (2016) 的 refund.shiny 软件包轻松可视化。

2.4 实践考量

应用快速广义函数主成分分析方法时，必须考虑若干实际问题。我们将在下文中详细讨论这些问题。

步骤 4 的必要性。步骤 4 之所以必要，是因为步骤 3 得到的特征函数估计量无偏，但受试者特异性随机效应的估计量有偏。这主要是由协方差算子的估计特征值存在偏差所致，而这种偏差随后会传递到得分上。图 1 展示了这种影响，其中绘制了步骤 3 和步骤 4 得到的估计得分与真实得分的对比（图 1(A)）、步骤 3 得到的估计得分与步骤 4 得到的估计得分的对比（图 1(B)），以及在对数优势尺度上步骤 3 与步骤 4 得到的估计曲线的对比。数据根据我们的模拟研究生成，属于二元函数型数据（详情见第4节），在这些数据中，fast-GFPCA 能够很好地估计真实特征函数。从图 1(A) 和图 1(B) 可以看出，步骤 3 得到的得分不仅存在偏差（远离恒等线，图 1(A) ），而且步骤 4 的得分实际上是对步骤 3 得分进行线性重缩放的结果（相关性近乎完美，但数据点偏离恒等线，图 1(B)）。这种偏差导致估计的对数优势与真实潜在函数的相关性大幅降低（图 1(C)）。请注意，图 1(B) 中的斜率几乎相同，这与第 2 节步骤2（关于分箱偏差的讨论）所示结果一致：特征值存在一个常数倍的偏差（因此得分也如此）。

可识别性。局部广义线性混合模型可能不可识别，或者模型拟合可能无法收敛。例如，对于某个分箱中的每位研究参与者，二元数据的观测值可能全部（或几乎全部）为 0 或 1。在这种情况下，局部模型不可识别。潜在的解决方案包括采用更大的分箱宽度、在线性预测尺度上施加下限或上限，或修改数据。例如，Agresti 和 Caffo（2000年）表明，在根据二元数据估计概率时，增加两个成功和两个失败的观测值可以改善置信区间的表现。在我们的研究情境中，可以通过在每个分箱中增加两个成功和两个失败的观测值来运用这一思路。在我们的应用和模拟中，增大箱尺寸就足以确保方法表现出色

方法。

分箱宽度。 如第 2 节所述，分箱宽度的选择是快速广义函数主成分分析算法中的一个重要调优参数。第4节中展示的模拟研究说明了这一点。我们需要在两个方面取得平衡：选择足够小的箱尺寸，以便估计潜在过程的曲率；同时又不能过小，以免导致广义线性混合模型拟合不稳定。一种可行的方法是逐步增大分箱大小，直至广义线性混合模型能够拟合数据。可以逐步增大分箱大小，并比较估计量的稳定性。另一种方法是在研究特定层面进行平滑，并检查绘图，以探索底层函数的复杂程度。

非重叠窗口与重叠窗口。使用非重叠窗口还是重叠窗口的选择因应用而异。非重叠窗口会减少需要估计的局部广义线性混合模型数量，但计算收益很小。相比之下，重叠窗口可以在更细的网格点s_{m_1}，……，s_{m_L}上进行估计，但可能会在估计的潜在过程中引入伪相关。这些自相关（如有）的影响目前在应用中尚不清楚，不过我们通过数据应用说明，该方法可能导致平滑不足。一个可能的技术解决方案是调整 FACE 的平滑选择准则，以允许处理自相关数据，但这超出了我们当前工作的范围。

2.5 替代方法

基于先行平滑个体曲线的所提方法至少有两种替代方案。首先，可以考虑先在响应尺度上对二元函数型数据进行平滑，并忽略数据的二元特性。随后，可使用链接函数（例如，针对二元数据使用逻辑函数）变换这些平滑后的数据，再对潜在尺度上的估计值应用函数主成分分析。另一种方案是通过拟合来进行平滑。

针对每个个体分别拟合广义加性模型 $g(E[Y_i(s)]) = \beta_{0i} + f_i(s)$，允许线性预测子在定义域上平滑变化，其中 $f_i(s)$ 使用丰富的基函数进行建模。随后，可以对所得的潜在尺度上的估计值应用函数主成分分析（fPCA）。Weishampel 等人（2023）最近提出了后一种方法。这两种方法都直观易懂，并且在某些情形下可能具有更快的计算速度。虽然它们可以用作快速的探索性工具，但我们将指出其中一些隐藏的缺陷。

考虑第一种方法。如果在响应尺度上进行平滑，并将 ~ ~平滑后的响应记为 $Y_i(s)$，那么一个基本要求是 $g\{Y_i(s)\}$ 必须有限且~有定义。例如，对于二元数据，需要满足 $Y_i(s) \in (0,1)$，并且最好与边界保持更远的距离，以确保 $\text{logit}\{\Pr(Y_i(s) = 1)\}]$ 有定义且表现良好。在零值很多的区域（例如身体活动数据中的夜间）和一值很多的区域（例如身体活动数据中的上午），这些平滑估计量分别会是 0 或 1。因此，需要借助人为设定的调优参数，将平滑估计值限制在远离 0 和 1 的范围内，再对其进行逆变换。这使得该方法依赖于个人选择，并且可能因界限参数的任意设定而引入无限偏差。

关于第二种方法，各个个体模型可能无法识别。事实上，考虑定义域中大片区域全为 0 或 1 的二元函数型数据。在这些区域中，广义加性模型只有依靠对系数函数 $f_i(s)$平滑性的假设才能实现识别，而这可能导致估计的对数优势发散。此外，所有平滑假设都是在个体层面作出的，没有考虑其他受试者提供的信息。在我们的数据应用中，许多参与者在清晨很早的时段（凌晨2点至5点）完全没有活动。相比之下，fast-GFPCA 方法会借助其他研究参与者的信息，在许多（但并非所有）研究参与者信息很少的区域，提供合理的受试者级别估计值。

除了上述提到的问题外，这两种方法都涉及调优参数。

与单个模型拟合的估计方法有关。例如，在惩罚样条的情境下，必须选择样条的数量、基函数以及平滑参数选择方法。鉴于生成模型意味着每个 $i = 1, \ldots, N$ 潜在函数的“波动程度”相同，后一项似乎尤为重要。分别拟合模型并不会强制满足这一约束，因此可能导致某些函数被估计为完全平滑，而另一些则相当起伏。此外，虽然 fast-GFPCA 能够处理数据集中的大量缺失区域，但本文描述的两种替代方法无法做到这一点，尤其是在存在过多的零或一时。

然而，这两种方法最重要的缺点是，我们无法成功地将它们用于 NHANES 应用；关于这些问题的详细讨论，请参见补充材料的 S-2 节。此外，fast-GFPCA 可以轻松扩展到协变量调整以及多层/结构化广义函数数据。目前尚不清楚上述两种方法是否都能直接扩展到这些情形。

3 应用

3.1 NHANES 加速度计数据

美国国家健康与营养调查（NHANES）是一项大型且持续开展的研究，提供了美国非机构化人群的全国代表性样本。NHANES 由疾病控制中心（CDC）开展，每两年进行一轮调查，旨在提供有关美国人口健康与营养状况的信息。NHANES 在 2003—2004年、2005—2006年、2011—2012年和 2013—2014年调查波次中部署了可穿戴加速度计。

NHANES研究2003—2006年的加速度测量部分要求参与者在清醒时段佩戴腰戴式加速度计。这是一份关于如何分析这些数据的指南。

相关内容见 Leroux 等（2019），数据包 `R` Leroux（2022）可在 GitHub 上通过 `https://github.com/andrew-leroux/rnhanesdata` 公开获取。2011—2014 年加速度计数据于 2021年 发布，提供多种分辨率：受试者级别、分钟级别和亚秒级别。受试者级别和分钟级别的数据基于新的监测器无关运动汇总（MIMS）单位，总结个体的加速度模式（John 等（2019））。本文使用 2011—2014 年分钟级别的 MIMS 数据，为参与者构建活动/不活动剖面。

为获得二元活动/不活动剖面，我们首先对参与者的每日 MIMS 数据进行阈值化，记为 $Y_{ih}^{B}(s) = 1\{Y_{ih}(s) \geq 10.558\}$，其中 $Y_{ih}(s)$ 对应于 i^{th} 个体在第 h 天第 s 分钟的 MIMS 单位。然后，我们将其活动/不活动剖面定义为 $Z_i(s)$ =中位数$\{Y_{ih}^{B}(s): h = 1, \ldots, H_i\}$。例如，如果 $H_i = 7$，那么当研究参与者 i 至少有 4 天在时刻 s 处于不活动状态时，$Z_i(s)$ 为 0，否则为 1。当良好天数为偶数时，例如 $\frac{H_i}{2}$，中位数定义为第 $\frac{H_i}{2} + 1^{\text{th}}$ 大的观测值。根据 Karas 等（2022）的建议，MIMS 单位尺度上活动/不活动的阈值设为 10。558，不过我们的方法同样适用于其他阈值。分析样本包含 $N = 4286$ 名参与者的数据，每人有 1440 个观测值（一天中的分钟数）。

尽管 NHANES 是一个全国代表性样本，但要获得针对总体数量和模型参数的全国代表性估计值，需要使用调查权重并考虑调查设计 Korn 和 Graubard（2011）；Lumley（2004）；Skinner 等（2017）。调查统计学与函数型数据分析的交叉领域是一个相对新兴的研究方向 Cardot 等（2013a、b，2014）；Parker 和 Holan（2022），超出了本研究的范围。因此，我们在数据应用中未考虑 NHANES 调查设计，不过这仍是未来方法学发展的一个重要方向。

3.2 比较方法与标准

我们使用分箱宽度（w_l）为 6、10 和 30 分钟的区间（包括重叠区间和非重叠区间），将 fast-GFPCA（标记为 `fastGFPCA`）应用于 NHANES 数据。我们将这些方法与 `registr::bfpca()` 函数中的快速二元变分函数主成分分析实现（标记为 `vbFPCA`）进行了比较 Wrobel（2018）；Wrobel 等人（2019）。我们还考虑了 fast-GFPCA 的一个修改后版本（标记为 `modified fastGFPCA`），该版本通过将模型 (2) 拟合到数据的四个子样本，进一步加快了步骤 4 中的`fastGFPCA`。`modifiedfastGFPCA` 方法详见补充材料的 S-3 节。Gertheiss 等人（2017）所述的方法在计算上无法应用于 NHANES 数据。为便于比较不同模型，我们将所有方法的主成分数量固定为 $K = 4$。

对于每种方法，我们比较模型参数和预测性能。在模型参数方面，我们比较前四个估计特征函数和总体均值函数。在预测性能方面，我们比较与每个模型拟合相关的估计样本内对数损失，以及估计的受试者工作特征曲线下面积（曲线下面积）。最后，我们比较各种方法的计算时间。尽管快速广义函数主成分分析的步骤2可以轻松并行化，但为给出该步骤所需时间的上限，我们报告的是串行拟合模型的计算时间。

使用 vbFPCA 方法和 fast-GFPCA 方法得到的总体均值估计值和特征函数存在显著差异（见第 3.3 节中的评论）。为探究这些差异，我们基于 NHANES 数据进行了简要的模拟研究，相关内容在补充材料的 S-4.1 节中进行了描述和总结。

3.3 结果

3.3.1 数据应用

ˆ**模型参数**。 图 2 展示了潜在过程的估计总体均值 $\beta_0(s)$（图 2A）和前四个特征函数 $\phi(s)$（图 2B）。图 2 的每一列对应一种不同的模型拟合，颜色表示所用方法（红色表示 vbFPCA，对应第 1–2 列；蓝色表示 fast-GFPCA，对应第 3–8 列）。修改版 fast-GFPCA 分别为随机选取的四个子样本估计总体均值函数，然后将这些估计值取平均，得到总体 $\beta_0(s)$。图 2A 第 3–8 列中的虚线展示了这些子样本的 $\beta_0(s)$ 估计值。由于修改版 fast-GFPCA 使用总体均值 $\phi(s)$，因此特征函数没有对应的虚线图。

对于修改版 fast-GFPCA 方法，总体均值在各子样本之间相当稳定（图 2A，第 3–8 列）。我们还发现，修改版与未修改版 fast-GFPCA 方法估计的线性预测子高度一致；参见图 3 ˆ ˆ 及相关讨论。此外，在不同的 fast-GFPCA 拟合结果中， $\beta_0(s)$ 和 $\phi(s)$ 相近（重叠窗口与非重叠窗口，以及分箱宽度），这表明 fast-GFPCA 算法对 NHANES 数据中输入参数的选择相当稳健。

然而，vbFPCA 与 fast-GFPCA 在估计的总体均值函数和前三个特征函数方面都存在显著差异。 $\beta_0(s)$ 的最大差异出现在早上 6 点至 8 点以及晚上 9 点至午夜前后；相较于 fast-GFPCA，vbFPCA 分别高估和低估了 $\beta_0(s)$。vbFPCA 的 $\beta_0(s)$ 表明，对于具有 $b_i(s) = 0$ 的参与者，早上 6 点至 8 点处于活动状态的概率约为 80%–90%，而 fast-GFPCA 给出的概率约为 50%。这一结果出乎意料，与数据中的结果不符；在数据中，我们观察到这段时间处于活动状态的概率约为 50%。一种可能的解释是，vbFPCA 给出的均值估计量存在偏差，且该偏差可能转移到了潜在随机过程变异中。我们的数据驱动模拟（见补充材料第 S-4.1 节）似乎支持这一假设。

目前尚不清楚导致 vbFPCA 方法出现这种偏差的原因，因为第 4 节所示模拟中并未出现相同情况，但这可能与我们的数据应用中夜间时段观测到的 0（非活动时段）比例较高有关。

线性预测子。 本节旨在了解不同方法估计的模型参数 $\beta_0(s)$ 和 $\phi(s)$ 之间的差异，如何导致估计的线性预测子 $\hat{}\eta_i(s)$ 出现差异；$i \in 1, ..., N$。为此，我们针对一天中的每一分钟，将 fast-GFPCA（采用非重叠窗口）得到的 $\hat{}\eta_i(s)$ 和 $w = 6$ 对其他四种 GFPCA 模型得到的 $\hat{}\eta_i(s)$ 进行回归。图3展示了这些时间特异性线性回归的结果。其形式为 $E[Y] = \mathcal{B}_0 + \mathcal{B}_1 X$，其中 Y 是 fast-GFPCA（采用无重叠窗口）得到的 $\hat{}\eta_i(s)$ 和 $w = 6$； X 是通过以下方法估计的 $\hat{}\eta_i(s)$：(1) fast-GFPCA，采用非重叠窗口和 $w = 30$ （橙色线）；(2) 修改版 fast-GFPCA，采用非重叠窗口和 $w = 6$ （蓝色线）；(3) vbFPCA，其中 Kt=8 （绿色线）；以及 (4) vbFPCA，其中 Kt=30$\hat{}\hat{}$（黄色线）。第一和第二个面板分别展示回归系数 $\mathcal{B}_0$ 和 $\mathcal{B}_1$，第三个面板展示方差解释百分比（R^2）。黑色虚线表示 Y 和 X 完全一致。

我们观察到，参考模型（无重叠的 fast-GFPCA 和 $w = 6$）与相应的修改版 fast-GFPCA 模型（蓝线；从左至右各面板的截距分别为 ≈ 0、斜率分别为 ≈ 1、$R^2 \approx 1$）之间几乎完全一致。类似地，我们观察到，跨越一天的、使用更大分箱宽度 $w = 30$ 的 fast-GFPCA 模型（橙线）之间存在近乎完美的线性关联（$R^2 \geq 0.95$，右侧面板），且在一天的活动时段（≈ 10上午至晚上8点）呈现近乎一比一的关系（截距 ≈ 0，斜率 ≈ 1）。在一天的活动时段内，vbFPCA 结果显示，均值也呈现类似的一比一趋势，但较低的 R^2 表明，fast-GFPCA 与 vbFPCA 方法之间的一致性低于使用不同输入参数得到的 fast-GFPCA 估计值之间的一致性。

预测准确性。 观察到的估计线性预测子差异，似乎并未转化为预测准确性的差异，无论是就曲线下面积还是对数损失而言；参见表1最右侧的两列。预测准确性相近但估计性能不同的一个可能解释是，模型之间预测结果的分歧主要发生在夜间时段（图3），此时个体处于活动状态的概率通常非常低。

计算时间。 表1中部列出了每种模型拟合的计算时间，其中上方各行为 fast-GFPCA 的结果，下方各行为 vbFPCA 方法的结果。对于 fast-GFPCA，计算时间按算法的各个步骤进行了细分。快速广义函数主成分分析算法（上方各行，“修改后” = “否”）需要约 3-4 小时，而 vbFPCA 在 Kt$=8$ 时需要 7 分钟，在 Kt$=30$ 时需要 20 分钟。这些计算时间主要由步骤 4 决定；如果模型拟合使用整个 NHANES 数据集，这一步不可避免。事实上，这里的 fast-GFPCA 方法是在保持数据函数结构的同时能够拟合的最简单广义线性混合模型。修改步骤 4 后（上方各行，“修改后” =“是”），计算时间大幅缩短，与 vbFPCA 方法相当。从实践角度来看，我们没有发现 fast-GFPCA 与修改版 fast-GFPCA 方法所得结果之间存在实质性差异。

4 模拟研究

我们的模拟旨在：(1) 量化随着样本量和网格大小增加，fast-GFPCA 的计算效率和可扩展性；(2) 将我们的方法与现有的广义函数主成分分析和二元函数主成分分析方法进行比较，评估其准确性；(3) 了解在第 2.2 节估计算法的步骤1中采用不同数据分箱策略时，我们的方法会呈现何种表现。对于较大的模拟数据集，我们重点关注二元函数型数据，因为唯一一种在计算上可行、能够处理大型函数型数据集的现有竞争方法，

专门针对二元数据设计。对于较小的数据集，我们还将我们的方法与一种现有的泊松函数数据广义函数主成分分析方法进行比较。

4.1 模拟设计

我们模拟二元函数型数据，取 $N = 100$、500、1000名受试者，在 $S = [0, 1]$ 上等间距网格中进行观测，该网格长度为$J = 100$、500、2000，并由所有受试者共享。由于竞争方法存在计算限制，泊松函数数据的模拟分别基于 $N = 50$、100和 $J = 100$、200。曲线取自模型（2），对应 $i \in 1$、.....、 N名受试者。我们构造 $K = 4$ 个主成分，其真实特征值为$\lambda_k = 0.5^{k-1}$； $k = 1$、2、3、4。真实特征函数取自两种旨在模拟真实数据场景的设定。在第一种设定中，潜在曲线 η_i（s）和特征函数具有周期性，其中 $\boldsymbol{\phi}$（s）$= \{\sqrt{2}\sin(2\pi s)\ \sqrt{2}\cos(2\pi s)\ \sqrt{2}\sin(4\pi s)\ \sqrt{2}\cos(4\pi s)\}$。第二种设定不具有周期性，真实特征函数为 $\boldsymbol{\phi}$（s）$= \{1, \sqrt{3}(2s-1)\ \sqrt{5}(6s^2-6s+1)\ \sqrt{7}($ $20s^3-30s^2+12s-1)\}$。在大多数模拟设定中，我们假设 β_0（s）$= 0$；不过，在部分场景中，我们使用B样条基构造非零的 β_0（s），具体针对以下二元函数型数据：（1） N、 $J =$（1000、2000）；（2） N、 $J =$（500、100）。

对于二元数据和泊松数据， $g(\cdot)$ 分别取为逻辑函数和对数链接。对于每个受试者和时间点，指数族观测 Z_i（s）独立地从概率为 $logit^{-1}[\eta_i(s)]$, 的伯努利分布，或速率为 $e^{[\eta_i(s)]}$的泊松分布中抽样。

4.2 与现有方法的比较

4.2.1 二元函数型数据

我们评估 fast-GFPCA 在不同分箱宽度 w_l 下的性能，并比较非重叠窗口和重叠窗口。具体而言，对于每种模拟场景，我们评估三种不同的分箱宽度 $w_l \in$（6,10, 50）。当 $J < 500$ 时，我们不使用 $w_l = 50$估计 fast-GFPCA，因为较大的箱尺寸在此背景下没有意义。

我们将 fast-GFPCA 与两种不同的二元函数主成分分析方法进行比较，这两种方法均通过 `registr` 包实现（Wrobel，2018；Wrobel 和 Bauer，2021）。第一种比较方法是 Gertheiss 等人（2017）提出的两步条件广义函数主成分分析模型，通过 `registr::gfpca twoStep()`函数实现，下文在正文和图中称为$tsGFPCA$ 。虽然 `gfpca twoStep()` 是一个通用函数，可处理多个指数族分布，但其计算量很大，因此在大多数模拟设置中并不实用。为降低计算负担，我们仅在 $N \in$ (100、500) 且 $J = 100$ 时实现 `tsGFPCA` 。在所有二元数据模拟设置中，我们都将 fast-GFPCA 与 Wrobel 等人（2019）的 vbFPCA 算法进行比较。该算法通过 `registr::bfpca()` 函数实现，记为 `vbFPCA`。这种方法计算效率很高，但专为采用 logit链接的二元函数型数据设计，无法推广到其他链接函数或指数族分布。由于 `vbFPCA` 方法使用 B 样条展开来建模总体均值 $\beta_0(s)$和特征函数 $\boldsymbol{\phi}(\boldsymbol{s})$ ，且未施加平滑惩罚，因此必须手动调整基函数的数量以获得最佳平滑度。为此，对于每个模拟数据集，我们采用 $Kt = 8$ 个基函数（软件包默认值）来实现 `vbFPCA` ，并使用 $Kt = 30$。

对于这两种竞争方法，我们均使用 `registr`选项 `periodic = TRUE`实现周期性 B 样条基。默认情况下，在 `registr` 包中，特征函数 $\boldsymbol{\phi}(\boldsymbol{s})$ 以未缩放的形式返回，且所在网格的大小为 100。为了与快速

广义函数主成分分析的结果进行比较，我们将使用$tsGFPCA$ 和 `vbFPCA` 估计的特征函数线性插值到大小为 J 的网格上，并按网格长度的平方根进行缩放。

4.2.2 泊松函数型数据

在泊松设定下，我们将 fast-GFPCA 与 `registr::gfpca twoStep()` 进行比较。由于比较方法的计算量很大，我们仅考虑样本量为 $N \in$ (50, 100)、网格长度为 $J \in$ (100,200) 的小规模数据设定。对于网格长度和样本量组合产生的六种模拟情景，我们分别模拟 100 个数据集。对于 fast-GFPCA，我们比较非重叠窗口和重叠窗口，并考虑分箱宽度 $w_l \in$ (6,10)。

4.3 评估标准

我们从以下方面比较三种方法（`tsGFPCA`、 `vbFPCA` 和 `fastGFPCA`）的表现：在预测变量线性空间 $\eta_i(s)$ 中恢复受试者特异性潜在均值的准确性、恢复总体均值 $\beta_0(s)$ 和特征函数$\boldsymbol{\phi}(\boldsymbol{s})$ 的准确性，以及计算效率。通过计算各模型中受试者对数优势比的平均积分平方误差（MISE）来量化受试者特异性对数优势比的估计准确性，定义为$\frac{1}{N}\sum_{i-1}^{N}\int_0^1(\hat{}\eta_i(s) - \eta_i(s))^2\,ds$。特征函数估计的准确性通过 MISE 进行比较，其定义为 $\frac{1}{k}\sum_{k=1}^{4}\int_0^1\left(\hat{}\phi_k(s) - \phi_k(s)\right)^2 ds$；总体均值的准确性则使用积分平方误差（ISE）衡量。计算时间以分钟为单位报告。

4.4 模拟结果：准确性

表 2、3 和 4 总结了各方法在不同模拟情景下关键量的准确性，其中结局数据来自二项式分布和泊松分布，且真实特征函数具有周期性。表 2 给出了 $\hat{}\eta_i(s)$ 的平均积分平方误差（MISE），即线性预测空间中估计的个体特异性潜在均值；对于二项式分布，这些均值为对数优势比。

泊松数据及对数率。同样，表 3 和表 4 分别总结了特征函数 $\phi(s)$ 的平均积分平方误差以及总体均值 $\beta_0(s)$ 的积分平方误差。

4.4.1 二项式数据

表 2 显示，所有方法都能准确估计潜在对数优势比。我们的 fastGFPCA方法与表现最佳的竞争方法 vbFPCA 使用 $Kt = 8$ 样条基底的效果同样出色，只有一种情景例外。模拟数据具有周期性，因此，采用重叠分箱的 fastGFPCA 优于采用非重叠分箱的 fastGFPCA 。除网格大小 J 最小（$J = 100$）时以外， fastGFPCA 在分箱宽度最大（ $w_l = 50$）时表现最佳。不过， fastGFPCA 在所有选定的分箱宽度下都表现良好。

表 3 表明，在每种情景下，我们的 fastGFPCA 方法恢复真实总体特征函数的表现都与竞争方法 vbFPCA 相当或更好。对于较小的网格大小 J，采用 $Kt = 8$的 vbFPCA 方法表现略好；但随着 J 增大， fastGFPCA优于 vbFPCA 。表 4 显示，在所有模拟情景下， fastGFPCA 恢复总体均值 β_0（s）的表现都优于 vbFPCA。这可能是因为 fastGFPCA 会对样条系数施加惩罚，从而在 β_0（s）的真实值附近得到平坦的估计曲线 $= 0$ ；而 vbFPCA不会对平滑性施加惩罚，因此当真实函数为线性时，无法同样准确地估计 β_0（s）。补充材料表 S1 中也观察到了非周期性数据的类似结果。

图4针对一个包含 $N = 500$ 个受试者和 $J = 500$ 个时间点、且总体均值函数 $\beta_0(t)$ 非零且为非线性的模拟情景，更细致地展示了这些结果。具体而言，图4展示了来自100个模拟数据集的估计总体均值函数，$\hat{}\hat{}\beta_0(s)$，以及前四个估计特征函数， $\phi_k(s)$; $k \in 1, \ldots, 4$。采用竞争性 vbFPCA方法的模型估计结果以红线表示，采用我们提出的 fastGFPCA 方法的模型估计结果以蓝线表示，黑色虚线表示真实值。在这一模拟情景下，所有方法都取得了合理的结果。我们采用重叠分箱且分箱宽度为 $w_l = 50$ 的 fastGFPCA 方法提供了

最佳结果。采用 $Kt = 8$ 样条基函数的 vbFPCA 方法表现也不错，不过在函数域 s的起始处高估了 β_0（t）。采用非重叠分箱的 fastGFPCA方法和 $w_l = 50$ 在函数域的端点处高估了 ϕ_4（s），这表明当数据非周期性且只能使用非重叠分箱进行 fastGFPCA时，较小的带宽可能更合适。

4.4.2 泊松数据

对于服从泊松分布的函数型数据，由于泊松函数主成分分析尚无变分 EM 方法，因此将所提出的 fastGFPCA方法与两步法 tsGFPCA进行了比较。表 3 显示，在每种模拟情景下，fastGFPCA 对潜在对数率的估计都远优于竞争方法 tsGFPCA 。同样，表 3 和表 4 显示，在所有情景下， fastGFPCA 对真实总体均值和特征函数的恢复都比 tsGFPCA 准确得多。在这些 fastGFPCA 方法中，采用重叠分箱的 fastGFPCA 和 $w_l = 10$ 通常表现最佳；但无论窗口是否重叠，也无论分箱宽度 w_l如何选择，不同的 fastGFPCA模型表现都相近。

补充材料中的图 S2 展示了一些直观解释，说明为何 tsGFPCA在泊松情形下表现不佳。该图显示了 25 个模拟数据集的估计总体均值函数和特征函数，每个数据集包含 $N = 100$ 名受试者和 $J = 100$ 个时间点。模型估计结果以红线（tsGFPCA）或蓝线（fastGFPCA）表示。所提出的 fastGFPCA 方法在分箱宽度为 $w_l = 6$和$w_l = 10$时表现合理，但 tsGFPCA 显然给出了错误的估计值。

4.5 仿真结果：计算效率

表 5 展示了二项式和泊松函数数据在各方法及模拟情景下的计算时间中位数（分钟）。在所有情景中，非重叠fastGFPCA 方法的计算效率都高于重叠 fastGFPCA 方法。分箱宽度 w_l 对非重叠分箱的计算时间影响可以忽略不计

，但对于重叠分箱，计算时间会随着分箱宽度增加而增加。`fastGFPCA` 在网格大小 J 增加时具有良好的扩展性，但随着受试者数量 N增加，扩展速度较慢。值得注意的是，在样本量较小时（$N \in \{100, 500\}$），`fastGFPCA` 的效率与 `vbFPCA`相当，甚至更高；`vbFPCA`是一种专为提高速度而设计的方法。对于 $N \geq 1000$，可以采用补充材料第 S-3 节中讨论的技术来加速`fastGFPCA` 。对于仅包含 100 名受试者、每名受试者有 200 个时间点的泊松数据，$tsGFPCA$ 的计算时间中位数为 92 分钟。这表明，对于我们这项包含 4286 名受试者、每人有 1440 个时间点的数据应用而言，$tsGFPCA$ 慢得难以接受。

5 讨论

本文提出的快速广义函数主成分分析方法，为非高斯数据的函数主成分分析这一复杂估计任务提供了一种简单、易懂且计算上可行的解决方案。此外，我们为快速广义函数主成分分析方法的基本原理提供了数学依据，并表明，与少数现有方法相比，该方法在估计精度和计算效率方面均具有优势。此外，本文介绍的用于二元数据的 vbFPCA 方法等现有方法，可能会对模型参数（总体均值函数和协方差算子）产生有偏估计。这表明，即使在某项应用中另一种方法的计算效率更高，快速广义函数主成分分析方法仍是一个合理的比较基准。

尽管本文表明，fast-GFPCA 速度快、准确，适合分析作为研究动机的 NHANES 数据，但方法学方面仍有待进一步研究。具体而言，目前尚不清楚如何选择最优分箱，包括如何确定分箱宽度，以及在估计时选择重叠窗口还是非重叠窗口。虽然基于交叉验证预测误差的准则可能是可行的选择，但在受试者层面进行交叉验证时，需要在非高斯模型中利用未纳入模型拟合的参与者数据预测随机效应，这并非易事

。此外，使用非重叠窗口时，在快速广义函数主成分分析算法的步骤 3 中，对潜在过程进行函数主成分分析时，自动选择平滑参数并不可靠。我们在此提出一种基于目视检查特征函数和/或估计协方差函数的临时解决方案。由于步骤 3 中实现的 FACE 方法速度很快，这种做法是可行的，但我们更希望有一种完全自动化的方法。针对 FACE 用于平滑参数选择的 GCV 准则，推导一种考虑自相关数据的适当变体，或许能改进本文提出的方法。尽管如此，本研究结果仍是向前迈出的令人鼓舞的一步，有助于在超高维数据中估计广义函数主成分分析，尤其是大型 N；这也是函数型数据分析方法在实践应用中的一个关键瓶颈。

fast-GFPCA 的一个吸引人之处在于，它可以扩展至：(1) 协变量依赖的广义函数主成分分析；(2) 多层次、结构化和纵向广义函数主成分分析。

协变量依赖的广义函数主成分分析。快速广义函数主成分分析方法可以轻松地将协变量纳入模型。考虑增加一个标量预测变量（例如年龄）的情况，记为 x_i。fast-GFPCA 的步骤2只需修改为拟合如下形式的局部模型$g(E[Z_i(s_j)|s_j \in S_l]) = \beta_0(s_{m_l}) + \beta_1(s_{m_l})x_i + b_i(s_{m_l}) = \eta_i(s_{m_l})$。随后，在步骤 4 中，最终的广义线性混合模型会纳入与所提出的变系数模型相关的加性项。进行这项扩展所需的工作极少。

嵌套、纵向或交叉设计的广义函数主成分分析。考虑每位研究参与者观测到多个函数的情况。例如，在 NHANES 数据中，每位研究参与者都有多天的加速度计数据。为简化符号，假设每位研究参与者都有 K 个函数。多层次广义函数主成分分析可自然地扩展快速广义函数主成分分析算法。具体而言，在步骤2中，我们拟合多层次模型 $g(E[Z_{ik}(s_j)]) = \beta_0(s_{m_l}) + b_i(s_{m_l}) + v_{ik}(s_{m_l})$。随后可进行步骤 3，

使用 MFPCA FACE Cui 等人 (2022年) 估计各层次的变异主方向。如果函数型数据采用纵向设计（Greven 等人，2010年）或交叉设计（Shou 等人，2015年），则可相应调整局部广义线性混合模型。

参考文献

Agresti, A. 和 Caffo, B.（2000）。通过增加两个成功和两个失败，得出简单有效的比例及比例差置信区间。*TheAmerican Statistician*, 54(4):280–288。Bates, D.、M¨achler, M.、Bolker, B. 和 Walker, S.（2015）。使用 lme4 拟合线性混合效应模型。统计软件杂志, 67(1):1–48。Cardot, H.、Dessertaine, A.、Goga, C.、Josserand, ´E. 和 Lardin, P.（2013a）。不同抽样设计的比较及用于估计函数型数据均值的置信带构建: 电力消耗案例分析。调查方法学, 39(2):283–301。Cardot, H.、Goga, C. 和 Lardin, P.（2013b）。均匀收敛以及抽样函数型数据均值的模型辅助估计量的渐近置信带。*Electronic journal of statistics*, 7:562–596。Cardot, H.、Goga, C. 和 Lardin, P.（2014）。高熵不等概率抽样设计下抽样函数型数据均值估计量的方差估计与渐近置信带。斯堪的纳维亚统计学杂志, 41(2):516–534。Chen, H.、Wang, Y.、Paik, M. C. 和 Choi, H. A.（2013）。一种用于降秩惩罚样条平滑的边际方法及其在多层函数型数据中的应用。美国统计学会杂志, 108(504):1216–1229。Chiou, J.-M.、Chen, Y.-T. 和 Yang, Y.-F.（2014）。多元函数主成分分析：一种归一化方法。*Statistica Sinica*, 24(4):1571–1596。Cui, E.、Li, R.、Crainiceanu, C. M. 和 Xiao, L.（2022）。快速多层函数主成分分析。计算与图形统计学杂志, 0(ja):1–33。Di, C.、Crainiceanu, C.、Caffo, B. 和 Punjabi, N.（2009）。多层函数主成分分析。应用统计年鉴, 3(1):458–488。Doherty, A.、Jackson, D.、Hammerla, N.、Pl¨otz, T.、Olivier, P.、Granat, M. H.、White, T.、van Hees, V. T.、Trenell, M. I.、Owen, C. G.、Preece, S. J.、Gillions, R.、Sheard, S.、Peakman, T.、Brage, S. 和 Wareham, N. J.（2017年）。使用腕戴式加速度计对大规模人群身体活动进行评估：UK Biobank 研究。*PLOS ONE*, 12(2):1–14。

Gertheiss, J.、Goldsmith, J. 和 Staicu, A.-M.（2017）。关于稀疏指数族函数响应曲线建模的说明。计算统计与数据分析，105：46–52。Goldsmith, J.、Scheipl, F.、Huang, L.、Wrobel, J.、Di, C.、Gellar, J.、Harezlak, J.、McLean, M.、Swihart, B.、Xiao, L.、Crainiceanu, C. 和 Reiss, P.（2020）。*refund*：函数数据回归。R软件包版本 0.1-23。Goldsmith, J.、Zipunnikov, V. 和 Schrack, J.（2015）。广义多层函数对标量回归与主成分分析。生物统计学，71(2)：344–353。Greven, S.、Crainiceanu, C.、Caffo, B. 和 Reich, D.（2010）。纵向函数主成分分析。*Electronic Journal of Statistics*，第 1022–1054 页。Hall, P.、M¨uller, H.-G. 和 Yao, F.（2008）。利用潜在高斯过程对稀疏广义纵向观测建模。皇家统计学会杂志：*B*辑（统计方法学），70(4)：703–723。Happ, C. 和 Greven, S.（2015）。针对不同（维度）定义域上观测数据的多元函数主成分分析。美国统计学会杂志，113：649 – 659。Hotelling, H.（1933年）。将统计变量复合体分解为主成分的分析。*Journal of Educational Psychology*，24：498–520。J.、P. 和 Paul, D.（2009）。一种基于几何的方法，用于从稀疏纵向数据中对函数主成分进行最大似然估计。*Journal of Computationaland Graphical Statistics*，18(4)：995–1015。James, G. M.、Hastie, T. J. 和 Sugar, C. A.（2000年）。稀疏函数型数据的主成分模型。*Biometrika*，87(3)：587–602。John, D.、Tang, Q.、Albinali, F. 和 Intille, S.（2019年）。用于加速度计数据处理的开源的与监测器无关的运动摘要。*Journal for the Measurement ofPhysical Behaviour*，2(4)：268 – 281。Jolliffe, I.（1982年）。关于回归中主成分使用的说明。*Journal of theRoyal Statistical Society, Series C*，31(3)：300–303。Jones, M. C. 和 Rice, J. A.（1992年）。展示大量相似曲线集合的重要特征。*The American Statistician*，46(2)：140–145。Karas, M.、Muschelli, J.、Leroux, A.、Urbanek, J. K.、Wanigatunga, A. A.、Bai, J.、Crainiceanu, C. M. 和 Schrack, J. A.（2022年）。基于加速度测量的身体活动指标比较：回顾性观察数据分析研究。*JMIRMhealth Uhealth*，10(7)：e38077。Korn, E. L. 和 Graubard, B. I.（2011年）。健康调查分析，第 323 卷。John Wiley & Sons。

Leroux, A. (2022). *rnhanesdata*：*NHANES* 加速度计数据处理流程。R 软件包版本 1.02。Leroux, A., Di, J., Smirnova, E., McGuffey, E. J., Cao, Q., Bayatmokhtari, E., Tabacu, L., Zipunnikov, V., Urbanek, J. K., and Crainiceanu, C. (2019). NHANES 中活动数据的整理与分析。生物统计学，11(2):262–287。Li, C. and Xiao, L. (2021). *mfaces*：多元稀疏函数型数据的快速协方差估计。R 软件包版本 0.1-3。Li, Z. and Wood, S. N. (2020). 使用离散化协变量加速大型广义线性模型的模型矩阵交叉乘积。统计与计算，30(1):19–25。Lumley, T. (2004). 复杂抽样调查样本分析。统计软件杂志，9(1):1–19。Parker, P. A. and Holan, S. H. (2022). 针对信息性抽样所收集调查数据的贝叶斯函数型数据模型及其在使用 NHANES 进行死亡率估计中的应用。生物统计学，n/a(n/a)。Pearson, K. (1901). LIII. 关于空间中点系的最佳拟合直线和平面。《伦敦、爱丁堡和都柏林哲学杂志与科学期刊》，2(11):559–572。Pinheiro, J. C. and Bates, D. M. (2000). *S* 和 *S-PLUS* 中的混合效应模型。施普林格，纽约。Ramsay, J. and Silverman, B. (2005). 函数型数据分析。施普林格纽约，纽约州，美国。Rice, J. and Silverman, B. (1991). 当数据为曲线时，以非参数方式估计均值与协方差结构。皇家统计学会杂志：*B*辑（方法论），53(1):233–243。Shou, H., Zipunnikov, V., Crainiceanu, C., and Greven, S. (2015). 结构化函数主成分分析。生物统计学，71(1):247–257。Skinner, C., Wakefield, J., et al. (2017). 复杂抽样调查数据设计与分析导论。统计科学，32(2):165–175。Staniswalis, J. and Lee, J. (1998). 纵向数据的非参数回归分析。美国统计学会杂志，93(444):1403–1418。van der Linde, A. (2008). 变分贝叶斯函数主成分分析。计算统计与数据分析，53(2):517–533。van der Linde, A. (2009). 一种针对二元和计数数据的函数主成分分析的贝叶斯潜变量方法。*AStA*统计分析进展，93(3):307–333。

Weishampel, A., Staicu, A.-M., and Rand, W. (2023). 使用广义函数型数据分析对社交媒体用户进行分类。计算统计与数据分析，179:107647。Wood, S. N. (2011). 半参数广义线性模型的快速稳定限制最大似然与边际似然估计。皇家统计学会杂志：*B*辑（统计方法学），73(1):3–36。Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. (2017). 超大规模数据的广义加性模型：英国黑烟网络每日数据建模。美国统计学会杂志，112(519):1199–1210。Wood, S. N., Scheipl, F., and Faraway, J. J. (2013). 混合模型中的简明中间秩张量积平滑。统计与计算，23(3):341–360。Wrobel, J. (2018). register：指数族函数型数据的配准。开放源代码软件杂志，3(22):557。Wrobel, J. and Bauer, A. (2021). registr 2.0：指数族函数型数据的不完整曲线配准。开放源代码软件杂志，6(61):2964。Wrobel, J., Park, S. Y., Staicu, A. M., and Goldsmith, J. (2016). 函数型数据分析的交互式图形。*Stat*，5(1):108–118。Wrobel, J., Zipunnikov, V., Schrack, J., and Goldsmith, J. (2019). 指数族函数型数据的配准。生物统计学，75(1):48–57。Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. (2018). 稀疏函数型数据的快速协方差估计。统计与计算，28(3):511–522。Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. (2021). *FACE*：稀疏函数型数据的快速协方差估计。R 软件包版本 0.1-6。Xiao, L., Zipunnikov, V., Ruppert, D., and Crainiceanu, C. (2016). 高维函数型数据的快速协方差估计。统计与计算，26(1):409–421。Yao, F., M¨uller, H., Clifford, A., Dueker, S., Follett, J., Lin, Y., Buchholz, B., and Vogel, J. (2003). 函数型主成分得分的收缩估计及其在血浆叶酸群体动力学中的应用。生物统计学，59(3):676–685。Yao, F., M¨uller, H.-G., and Wang, J.-L. (2005). 稀疏纵向数据的函数型数据分析。美国统计学会杂志，28(100):577–590。

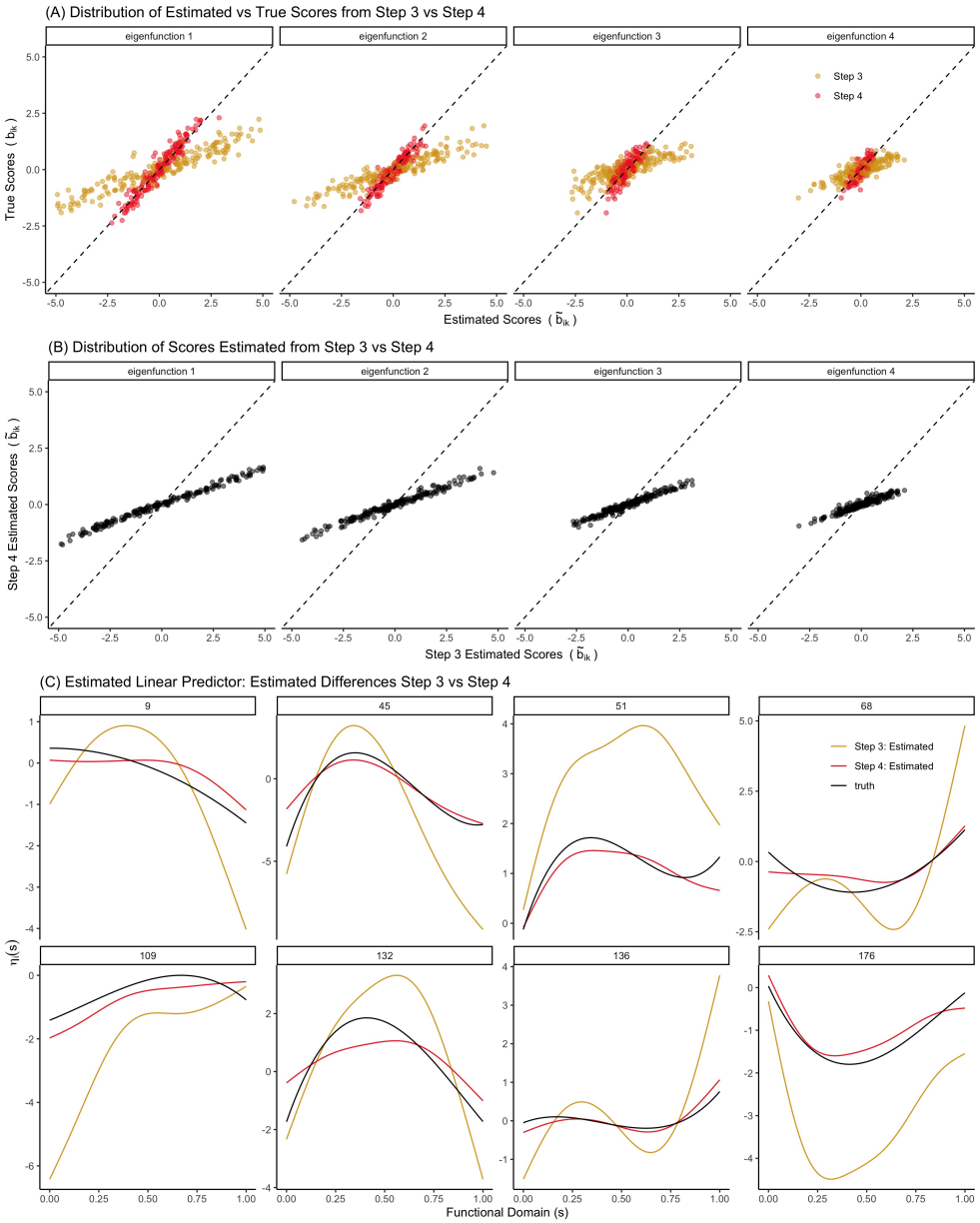


图 1：使用模拟研究中的一个模拟数据集（$N = 200$, $J = 200$, $w_l = 10$, $\phi(s) = \{\sqrt{2}\sin(2\pi s)\ \sqrt{2}\cos(2\pi s)\ \sqrt{2}\sin(4\pi s)\ \sqrt{2}\cos(4\pi s)\}$）说明快速广义函数主成分分析方法中步骤 4 的必要性。(A) 分别针对前四个特征函数，绘制步骤 3（金色点）和步骤 4（红色点）估计得分（x 轴）与真实得分（y 轴）的关系。黑线表示恒等线。(B) 绘制步骤 3 估计的得分（x 轴）与步骤 4 估计的得分（y 轴）的关系。(C) 线性预测尺度上的估计曲线



图 2：NHANES 中估计的总体均值函数（第一行）和前四个估计特征函数（第 $2-5$ 行）。模型估计结果以红线（vbFPCA）或蓝线（fast-GFPCA）表示。最左侧两列对应 vbFPCA（第 1 列中的 Kt=8 和第 2 列中的 Kt=30 ）。最右侧六列对应使用不同输入参数（重叠窗口与非重叠窗口）和窗口大小（$w = 6$, 10，30）的 fast-GFPCA。基于修改版 fast-GFPCA 得到的总体均值函数估计值以虚线表示。

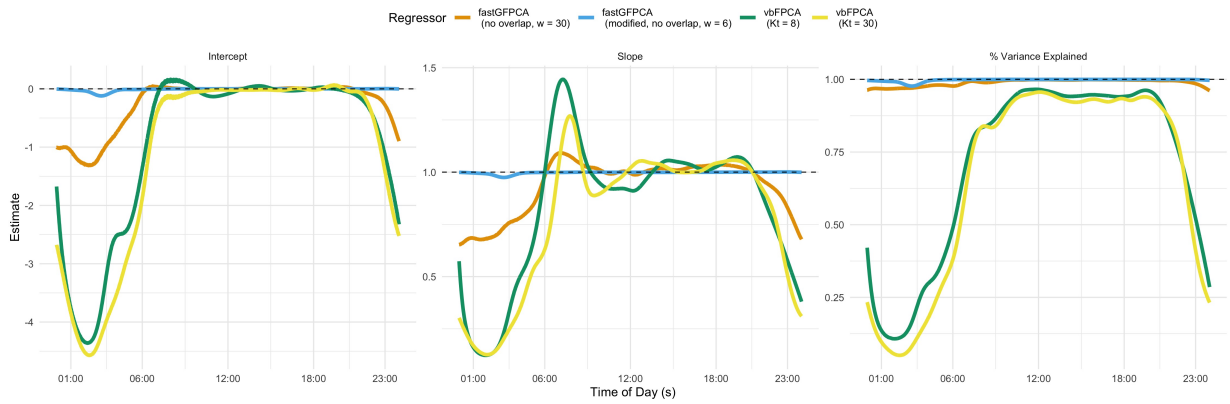
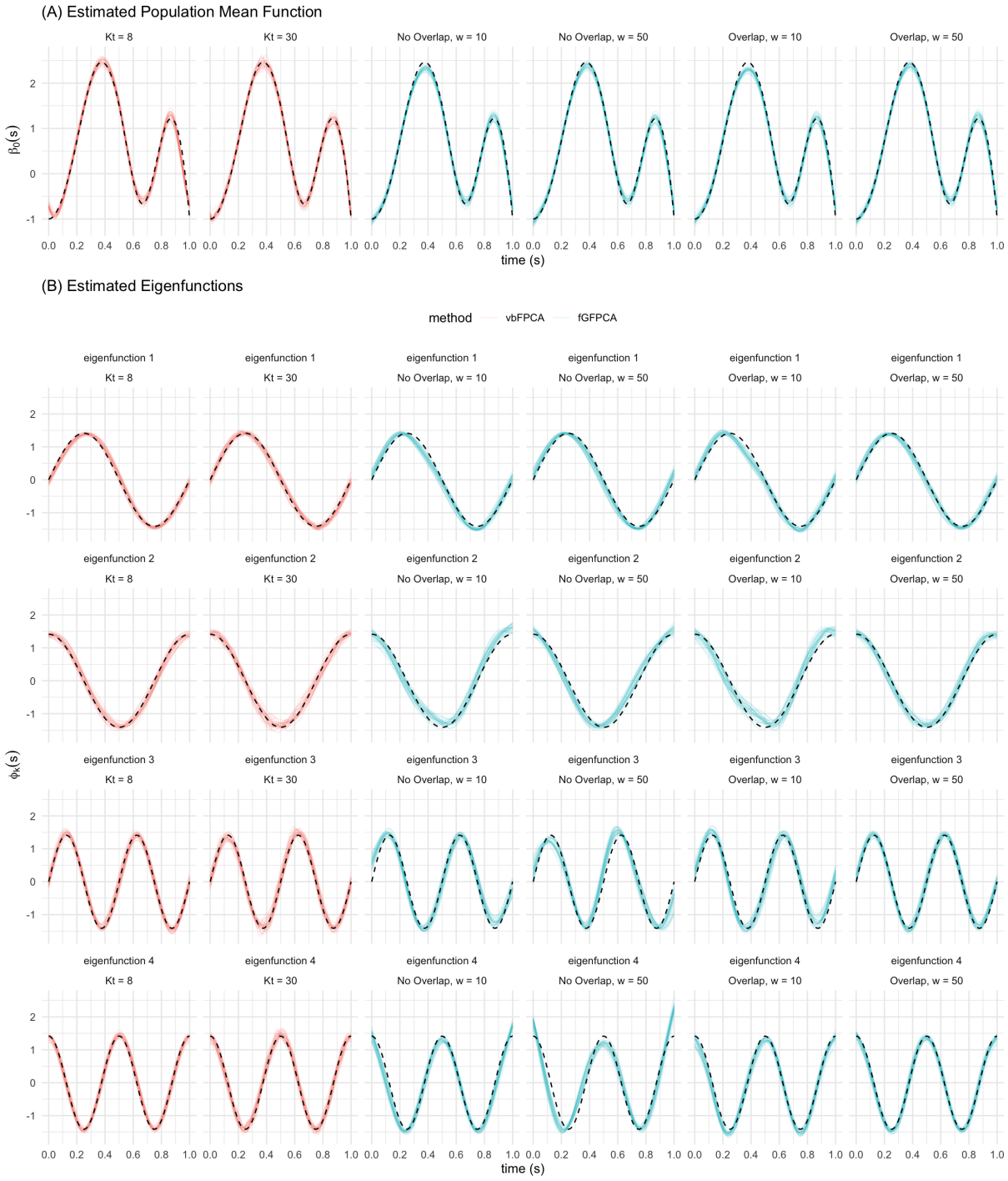


图3：使用 $w_l = 6$ 和非重叠窗口，通过 fast-GFPCA 得到的活动状态对数优势估计值进行时间特异性回归的结果，以及：(1) 使用 $w_l = 30$ 和非重叠窗口，通过 fast-GFPCA 得到的结果（红线）；(2) Kt $= 8$ 的 vbFPCA 结果（绿线）；(3) Kt =30 的 vbFPCA 结果（蓝线）。对每分钟分别拟合形式为 $E[Y] = \mathcal{B}_0 + \mathcal{B}_1 X$ 的回归模型，并在各面板中（从左到右）分别绘制所得估计值 $\hat{\mathcal{B}}_0$、$\hat{\mathcal{B}}_1$ 和方差解释百分比（R^2）。各面板中与黑色虚线的偏离表示回归变量与使用 $w_l = 6$ 和非重叠窗口通过 fast-GFPCA 得到的结果之间的一致程度降低。

fastGFPCA								
参数	修改后	计算时间（分钟）					AUC	对数损失
		步骤1	步骤2	步骤 3	步骤 4	总计		
重叠，$w = 6$	No		16.42	0.04	174.20	190.63	0.909	0.328
	Yes		16.42	0.04	10.90	27.33	0.909	0.328
重叠，$w = 10$	No		23.28	0.04	184.82	208.12	0.909	0.327
	Yes		23.28	0.04	10.74	34.03	0.909	0.327
重叠，$w = 30$	No		55.90	0.04	184.66	240.57	0.909	0.327
	Yes		55.90	0.04	10.48	66.38	0.909	0.327
无重叠，$w = 6$	No		2.48	0.01	188.90	191.43	0.909	0.328
	Yes		2.48	0.01	11.34	13.13	0.909	0.328
无重叠，$w = 10$	No		2.19	0.01	185.04	187.28	0.909	0.327
	Yes		2.19	0.01	10.89	13.13	0.909	0.327
无重叠，$w = 30$	No		1.97	0.01	197.35	199.36	0.909	0.327
	Yes		1.97	0.01	11.60	13.61	0.909	0.327
vbFPCA (registr::bfpca())								
参数		计算时间（分钟）					AUC	对数损失
Kt=8						7.11	0.909	0.328
Kt=30						20.86	0.910	0.326

表1：fastGFPCA（上方各行）在不同参数设置下（第1列：重叠窗口与非重叠窗口，窗口大小为 $w = 6$, 10，30），分别使用主要算法和修改算法（第2列：修改后 =否或是）所得的计算时间及样本内预测性能汇总；以及采用 $Kt = 8$ 或 $Kt = 30$估计的变分贝叶斯方法（vbFPCA，下方各行）。对于fastGFPCA，计算时间同时列出总时间和过程各步骤的用时（步骤1留空，因为从计算角度看该步骤实际上是瞬时完成的）。最右侧两列所示的曲线下面积和对数损失，是基于一天中的所有分钟计算得出的。



^^图4：从包含 $N = 500$ 名受试者和 $J = 500$ 个时间点的100个模拟二元函数型数据集中得到的估计的总体均值函数 $\beta_0(s)$和特征函数 $\phi(s)$。模型估计值以红线（vbFPCA）或蓝线（fastGFPCA）表示。将使用 $Kt = 8$ 个（左列）或$Kt = 30$ 个（第二列）基函数的模型所得到的vbFPCA估计值，与六个fastGFPCA模型进行比较。从左到右，fastGFPCA模型（蓝色列）分别使用分箱大小为 $w_l = 10$和 $w_l = 50$的非重叠窗口，以及分箱大小为 $w_l = 10$ 和 $w_l = 50$的重叠窗口进行估计。

$\eta_i(s)$的平均积分平方误差											
			fast-GFPCA						vbFPCA		tsGFPCA
			重叠分箱			非重叠分箱					
族	N	J	$w_l = 6$	$w_l = 10$	$w_l = 50$	$w_l = 6$	$w_l = 10$	$w_l = 50$	K $= 8$	K $= 30$	K $= 30$
二项式	100	100	2.19	**2.14**	-	2.39	2.44	-	2.15	2.87	2.18
		500	0.55	0.52	**0.49**	0.53	0.53	0.58	**0.49**	0.63	-
		2000	0.17	0.15	**0.13**	0.15	0.15	**0.13**	**0.13**	0.16	-
	500	100	2.05	**2.02**	-	2.33	2.43	-	**2.02**	2.15	2.03
		500	0.5	0.49	**0.47**	0.5	0.5	0.58	**0.47**	0.5	-
		2000	0.15	0.14	**0.12**	0.14	0.14	0.13	**0.12**	0.13	-
	1000	100	2.05	2.04	-	2.33	2.41	-	**2.01**	2.08	-
		500	0.5	0.48	**0.47**	0.51	0.5	0.58	**0.47**	0.48	-
		2000	0.14	0.13	**0.12**	0.14	0.13	0.13	**0.12**	**0.12**	-
泊松	50	100	0.52	**0.49**	-	0.56	0.65	-	-	-	13.0
		200	0.32	**0.28**	-	0.32	0.29	-	-	-	15.0
	100	100	0.45	**0.43**	-	0.5	0.59	-	-	-	15.9
		200	0.25	**0.24**	-	0.27	0.25	-	-	-	20.5

表 2：各方法和模拟情景下，线性预测空间中估计的个体特异性潜在均值 $\hat{\eta}_i(s)$ 的平均积分平方误差（MISE）。每行中，该模拟情景下 MISE 最低的方法以**粗体**显示。本表汇总的所有数据均通过总体均值 $\beta_0(s) = 0$ 模拟得到。“-”表示该模型未针对相应模拟情景进行评估。所有数值均乘以 10。

特征函数的平均积分平方误差 $\phi(s)$											
			fast-GFPCA						vbFPCA		tsGFPCA
			重叠分箱			非重叠分箱					
族	N	J	$w_l = 6$	$w_l = 10$	$w_l = 50$	$w_l = 6$	$w_l = 10$	$w_l = 50$	K $= 8$	K $= 30$	K $= 30$
二项式	100	100	0.91	0.66	-	1.13	1.56	-	**0.47**	1.13	0.48
		500	0.66	0.55	0.46	0.57	0.59	0.58	**0.28**	0.39	-
		2000	0.64	0.54	**0.42**	0.55	0.52	**0.42**	0.72	0.88	-
	500	100	0.19	0.18	-	0.58	0.8	-	**0.12**	0.26	0.13
		500	0.13	0.11	**0.08**	0.13	0.11	0.32	**0.08**	0.11	-
		2000	0.14	0.12	**0.1**	0.13	0.13	**0.1**	0.56	0.62	-
	1000	100	0.11	0.08	-	0.5	0.7	-	**0.05**	0.11	-
		500	0.06	0.04	**0.03**	0.07	0.07	0.26	0.05	0.07	-
		2000	0.05	0.04	**0.03**	0.05	0.04	0.04	0.52	0.66	-
泊松	50	100	1.03	**0.94**	-	1.16	1.24	-	-	-	10.97
		200	0.96	**0.88**	-	0.91	0.98	-	-	-	8.8
	100	100	0.4	**0.34**	-	0.52	0.66	-	-	-	10.79
		200	0.37	**0.34**	-	0.4	0.41	-	-	-	8.7

表 3：估计的总体层面潜在特征函数的平均积分平方误差（MISE），$\phi_k(s)$；$k = 1$、.....、4。每一行中，在该模拟场景下 MISE 最低的方法以**粗体**显示。本表汇总的所有数据均使用总体均值 $\beta_0(s) = 0$ 进行模拟。“-”表示该模型未针对相应模拟场景进行评估。所有数值均乘以 10。

总体均值 $\beta_0(s)$ 的积分平方误差											
			fast-GFPCA						vbFPCA		tsGFPCA
			重叠分箱			非重叠分箱					
族	N	J	$w_l = 6$	$w_l = 10$	$w_l = 50$	$w_l = 6$	$w_l = 10$	$w_l = 50$	K = 8	K = 30	K = 30
二项式	100	100	**1.34**	1.35	-	2.42	1.5	-	15.69	27.37	1.85
		500	0.53	0.54	0.48	**0.42**	**0.42**	0.62	12.94	15.88	-
		2000	0.1	0.09	0.11	**0.06**	0.08	0.07	12.13	12.11	-
	500	100	0.19	**0.17**	-	0.3	0.41	-	3.53	5.74	0.46
		500	0.06	0.06	0.06	**0.05**	**0.05**	0.06	3.03	3.45	-
		2000	0.02	0.02	**0.01**	0.02	0.02	0.02	3.12	3.13	-
	1000	100	0.1	**0.09**	-	0.14	0.12	-	1.69	2.87	-
		500	0.04	0.04	0.04	**0.03**	0.05	0.04	1.57	1.77	-
		2000	0.02	0.02	0.02	0.02	0.02	**0.01**	2.06	1.87	-
泊松	50	100	4.34	**3.46**	-	5.12	5.54	-	-	-	116.59
		200	2.43	**2.18**	-	3.68	3.45	-	-	-	126.52
	100	100	2.31	**2.18**	-	3.27	4.81	-	-	-	174.8
		200	**1.64**	1.67	-	4.1	2.36	-	-	-	148.47

表 4：估计的总体层面潜在均值 $\beta_0(s)$ 的积分平方误差（ISE）。在每一行中，该模拟场景下 ISE 最低的方法以**粗体**标示。本表汇总的所有数据均使用总体均值 $\beta_0(s) = 0$ 进行模拟。“-”表示该模型未针对相应的模拟场景进行评估。所有数值均乘以 10^3 倍。

中位数计算时间											
			fast-GFPCA						vbFPCA		tsGFPCA
			重叠分箱			非重叠分箱					
族	N	J	$w_l = 6$	$w_l = 10$	$w_l = 50$	$w_l = 6$	$w_l = 10$	$w_l = 50$	K = 8	K = 30	K = 30
二项式	100	100	0.1	0.1	-	0.1	**0**	-	**0**	0.2	16.9
		500	0.5	0.5	1.6	**0.1**	**0.1**	**0.1**	**0.1**	0.7	-
		2000	2	2.2	6.5	0.3	**0.2**	**0.2**	0.9	4.7	-
	500	100	2.3	2.3	-	2.2	2.2	-	**0.1**	0.7	71.3
		500	2.7	3.2	7	2	1.9	1.9	**0.6**	2.8	-
		2000	7.4	9.5	27	**3.1**	**3.1**	**3.1**	5.7	24.4	-
	1000	100	20.4	20.1	-	20.2	20.4	-	**0.4**	1.3	-
		500	18.6	20	29.9	17.2	17.4	17.2	**1.7**	6.4	-
		2000	30.1	35.1	75.9	22.3	22.6	22.4	**21**	71.4	-
泊松	50	100	0.1	0.1	-	**0.02**	**0.02**	-	-	-	11.2
		200	0.2	0.21	-	0.04	**0.03**	-	-	-	40.4
	100	100	0.15	0.16	-	**0.06**	**0.06**	-	-	-	34.6
		200	0.25	0.27	-	0.07	**0.06**	-	-	-	92.0

表 5：不同方法和模拟情景下计算时间（分钟）的中位数。每行中，该模拟情景下计算时间最快的方法以**粗体**显示。本表汇总的所有数据均是在总体均值为 $\beta_0(s) = 0$ 的情况下模拟得到的。“-”表示该模型未针对相应模拟情景进行评估。

补充材料

S-1 快速广义函数主成分分析算法描述

算法1： Fast-GFPCA

步骤1：数据分箱

输入 ： $[\{Z_i(s_j)\}, 1 \le i \le N, 1 \le j \le J]$

输出： $[\{Z_i(s_j)\ \ l\} 1 \le i \le N, j \in \mathcal{S}_l, 1 \le l \le L\}]$

步骤2：估计局部广义线性混合模型

输入 ： $[\{Z_i(s_j)\ \ l\} 1 \le i \le N, j \in \mathcal{S}_l, 1 \le l \le L\}]$

输出： $[\{\hat{}\eta_i(s_{m_l})\ \ l\}, 1 \le i \le N, 1 \le l \le L]$

对于 $l = 1, \ldots, L$ **do**

拟合以下形式的广义线性混合模型:

$$g(E[Z_i(s_j)|s_j \in S_l]) = \beta_0(s_{m_l}) + b_i(s_{m_l}) = \eta_i(s_{m_l})$$

获取 $\widehat{\eta}_i(s_{m_l}) = \hat{\beta}_0(s_{m_l}) + \widehat{b_i(s_{m_l})}$

end

步骤 3：对局部潜变量估计值进行函数主成分分析

输入 ： $[\{\widehat{\eta}_i(s_{m_l})\ \ l\}, 1 \le i \le N, 1 \le l \le L]$

输出： $[\{\quad \hat{\phi}_k(s_{m_l})\ \ l\}, \quad 1 \le l \le L, \quad 1 \le k \le K]$

使用 FACE算法估计 $\{\widehat{\eta}_i$ 的协方差函数 $\mathrm{Cov}\{\widehat{\eta}_i(u)\ \widehat{\eta}_i(v)\}$ ，其中 $u, \quad v = m_1, \quad \ldots, \quad m_L$

获取 $\hat{\phi}_k$ （s） 对于 $s = m_1, \ldots, m_L$

步骤 4：重新估计个体特异性得分

输入 ：[{<escape>$\hat{\phi}_k$</escape>(<escape>s_l</escape>)<escape> $l\}, 1 \le l \le L, 1 \le k \le K\}$,[{$Z_i$ </escape>(<escape>s_j</escape>)<escape>}, $1 \le i \le N, 1 \le j \le J$]</escape>

输出： [{ $\widehat{b_i(s_j)}$, <escape>$\hat{\beta}_0(s_j)\}$</escape>, <escape>$1 \le i \le N$ </escape>, <escape>$1 \le j \le J$]</escape>

对于 $k = 1, \cdots, K$ **do**

获取 $\widehat{\quad}$在原始网格 $s_1, \ldots\ldots, s_J$ 上的 $\phi_k(s)$

end

使用以下形式的加性广义线性混合模型估计广义函数主成分分析

$$\sim g(E[Z_i(s)]|\{\phi_k(s) : 1 \le s \le J, 1 \le k \le K\}) = \beta_0(s) + b_i(s) = \beta_0(s) + \sum_{k=1}^{K}{}_{(s)} \sim \xi_{ik}\phi_k$$

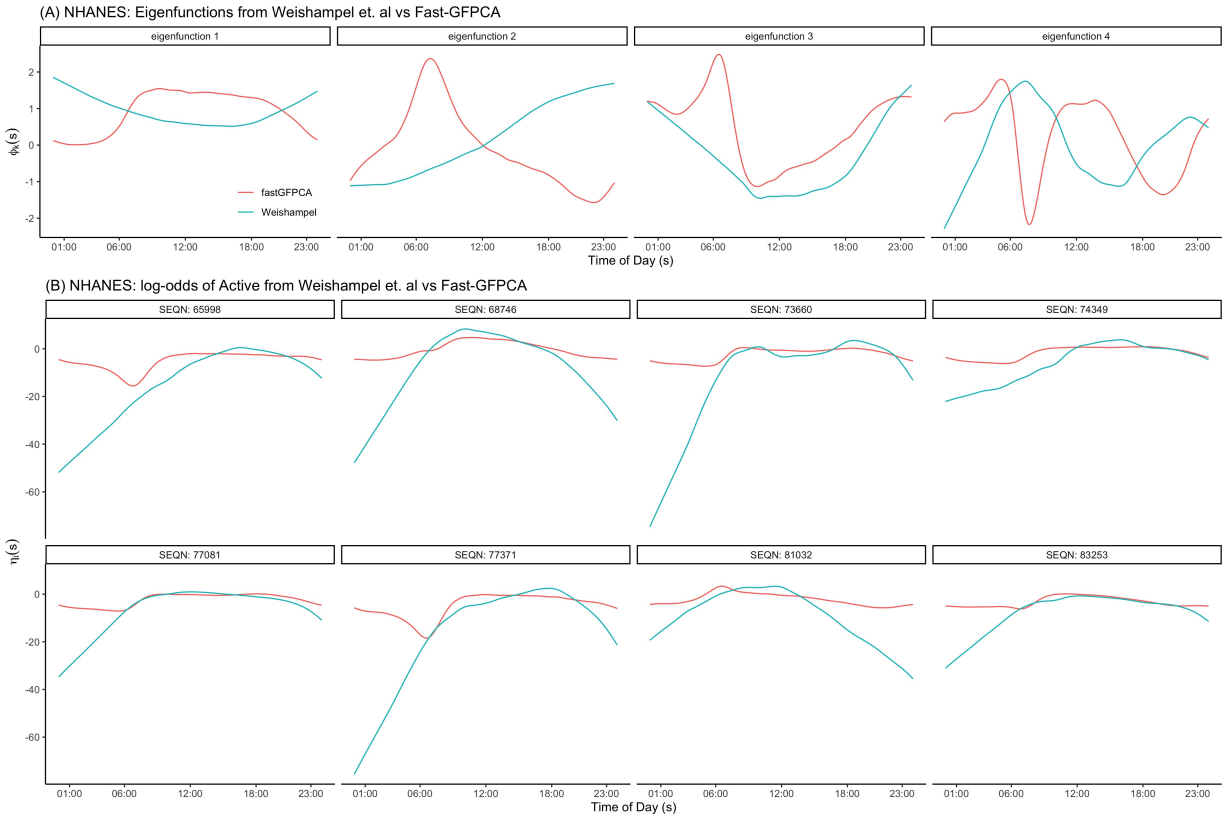
ξ_{ik} 独立同分布$\sim N(0, \sigma_k^2)$, $\quad$ 协方差$(\xi_{ik}, \ \xi_{il}) = 0$ 对于 $k \ne l$

S-2 与 NHANES 中的 Weishampel 进行比较

图 S1 展示了 Weishampel 等人（2023）提出的 GFPCA 方法与 fast-GFPCA 方法应用于 NHANES 数据后的结果。我们通过首先使用 30 个惩罚三次回归样条拟合个体模型，并采用 REML 进行平滑参数选择，实现了 Weishampel 等人（2023）提出的方法。在潜在尺度上进行的 FPCA 使用 `refund::fpca.face()` 实现。这些特征函数的形状存在显著差异（图 S1(A)），其中 Weishampel 等人（2023）的第一特征函数（图 S1(A)，左侧面板，蓝线）主要受函数两端的行为（清晨和深夜）影响。这可能是因为许多个体在清晨（以及在一定程度上的深夜）始终处于全 0（不活动）状态。在使用这些数据拟合个体 GAM 以估计参与者的潜在函数时，这往往会导致边界处的估计值出现发散。这种效应可以从图 S1(B) 中 Weishampel 等人（2023）方法得到的活动状态对数优势估计值中看出。图中将 Weishampel 等人（2023）的方法与 fast-GFPCA 的结果进行了比较，展示了随机选取的 8 名参与者（每个面板对应一名参与者），颜色表示所采用的估计方法（蓝色表示 Weishampel 等人（2023）的方法，红色表示 fast-GFPCA）。Weishampel 等人（2023）方法得到的每个估计值都在边界处出现发散，但发散程度各不相同。

S-3 进一步加快 fast-GFPCA

快速GFPCA的主要潜在计算瓶颈在于，需要在快速广义函数主成分分析算法的步骤 4 中重新估计个体特异性得分。我们已经确定了三种可能加快当前算法运行速度的途径，其中第一种应用于我们的数据应用中；至于每种途径的详细研究，则留待未来开展。第一种方法是在总体上估计步骤 1-3，以获得用于估计个体特异性随机效应的总体特征函数，但仅在随机抽取的数据



ˆ 图 S1：（A）fast-GFPCA 得到的前四个估计特征函数（$\phi_k(s)$）（红线）与 Weishampel 等人得到的结果（蓝线）的比较。（B）处于活跃状态的估计对数优势比。模型估计值以红线（fastGFPCA）或蓝线（fastGFPCA）表示。将 fastGFPCA 的估计值与两个 fastGFPCA 模型进行比较。fastGFPCA 模型（蓝色柱形）使用分箱大小为 $w_l = 6$ 和 $w_l = 10$的重叠窗口进行估计。

子样本上估计步骤 4。在步骤 4 中，所做的只是重新估计总体均值函数，并将个体特异性潜在估计投影到总体基函数估计上。对于较大的样本量，例如我们的数据应用，这两个量在足够大的子样本中都应相对稳定（例如 $N = 1000$）。

剩余两种方法基于我们在模拟和数据应用中的观察：步骤 3 中获得的得分与步骤 4 中获得的得分几乎完全相关，但两者之间存在一个线性缩放因子的差异（见图 1(B)）。对于每个特征函数，该缩放因子似乎相同或数值相近（注意，图 1(B) 中各特征函数对应的斜率几乎完全相同）。这表明，如果能够估计出该缩放因子，就可以得到受试者特异性得分的最终估计值。

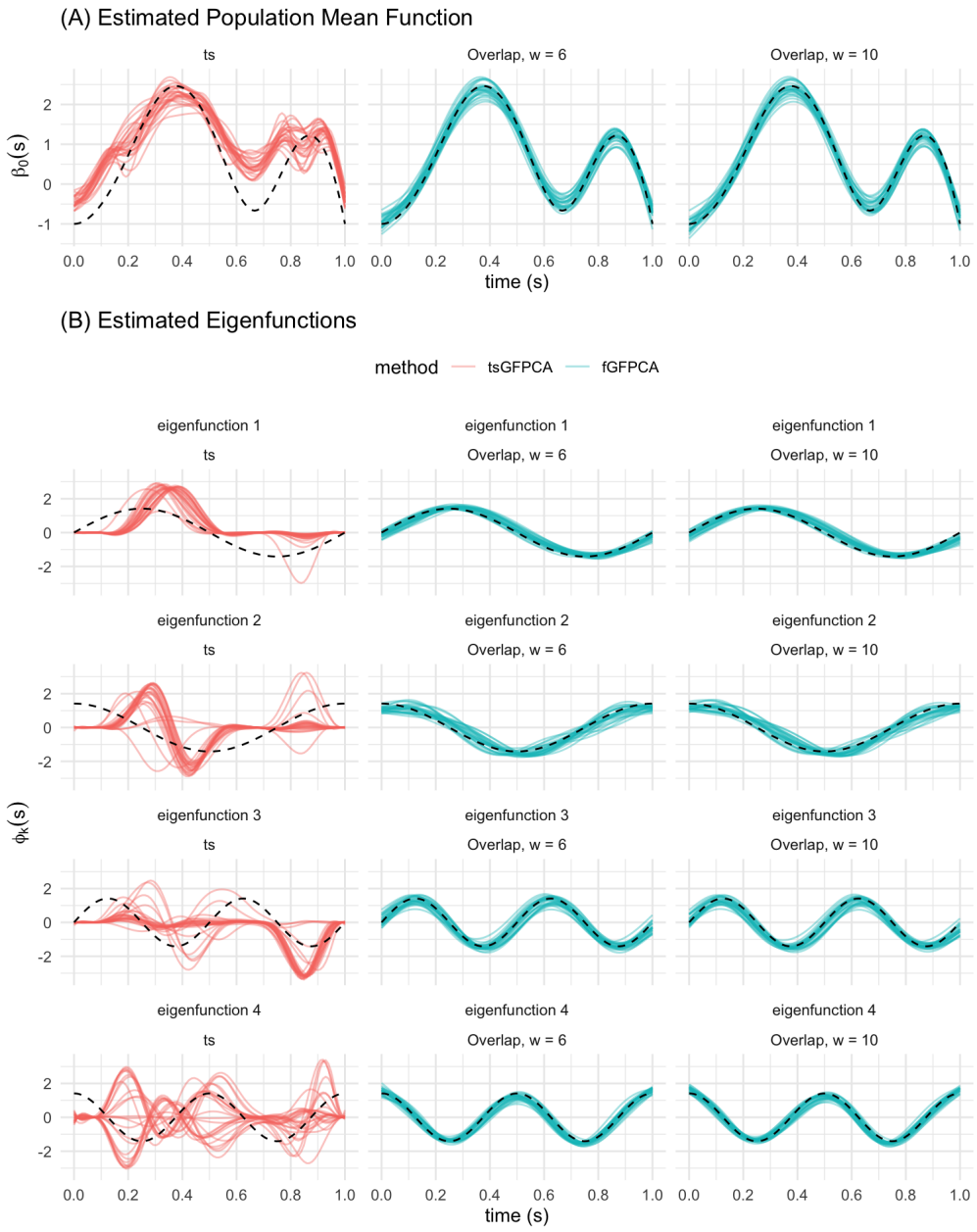
所需的计算量会更少。我们认为至少可以通过两种方式实现这一点。首先，可以基于数据子集估计步骤 4，然后将缩放因子直接应用于通过步骤 3 得到的估计值。或者，可以采用先前建议的子集分析来估计个体得分方差 λ_k，然后将估计特征函数（$[\{\phi_k(s)\}, 1 \leq s \leq J]$）和方差参数都视为固定值。这样，步骤 4 中的模型只需估计总体均值（$\beta_0(s)$），从而大幅降低步骤 4 的计算复杂度。这可以通过 $mgcv$ 软件包轻松实现，我们在数据应用中展示了这一思路，但对该思路的严格考察留待未来工作。

S-4 附加模拟结果

S-4.1 数据驱动模拟

为探究 vbFPCA 与 fast-GFPCA 方法在均值和主成分估计方面的差异，我们进行了一项简短的模拟研究。首先，使用窗口大小为 $w = 6$ 的非重叠 fast-GFPCA 模型，为 $K = 4$ 个特征函数估计模型 (2) 的参数。我们生成了五个数据集，每个数据集包含 $N = 4000$ 位参与者，且均在与原始数据相同的网格 ($J = 1440$) 上进行观测。由于我们发现，在 fast-GFPCA 方法中，参数估计值不会因分箱宽度以及采用重叠还是非重叠窗口而有所不同，因此我们仅使用 $w = 6$ 和非重叠窗口来估计 fast-GFPCA 模型。

图 S3 展示了这项小规模模拟研究的结果，其结构与图 2 相同，估计的总体均值函数 $\hat{\mu}(s)$ 和特征函数 $\phi_k(s)$ 分别绘制在第一行和第二至第五行。真实的均值和特征函数以黑色实线表示，而 5 个模拟数据集的估计值则以半透明的红色（vbFPCA）和蓝色（fast-GFPCA）模型表示，



ˆ 图 S2：来自 25 个模拟的泊松函数数据集的估计总体均值函数 ($\beta_0(s)$) 和前四个估计特征函数 ˆ($\phi_k(s)$)，数据集包含 $N = 100$ 个受试者和 $J = 100$ 个时间点。模型估计以红线（fastGFPCA）或蓝线（fastGFPCA）表示。将 fastGFPCA 估计值与两个 fastGFPCA 模型进行比较。fastGFPCA 模型（蓝色列）使用分箱大小为 $w_l = 6$ 和 $w_l = 10$ 的重叠窗口进行估计。

分别。在 vbFPCA 方法中，总体均值函数和特征函数的估计值均存在偏差，且该偏差与我们从真实数据中得到的估计值高度吻合。相比之下，fast-GFPCA 估计值的偏差极小。由于

正如预期的那样，在样本量相对较大时（$N = 4000$），总体均值和特征函数的估计值在各个模拟数据集中都极为一致。这些发现表明，在我们的数据应用中，vbFPCA 方法确实可能给出了有偏的参数估计值。

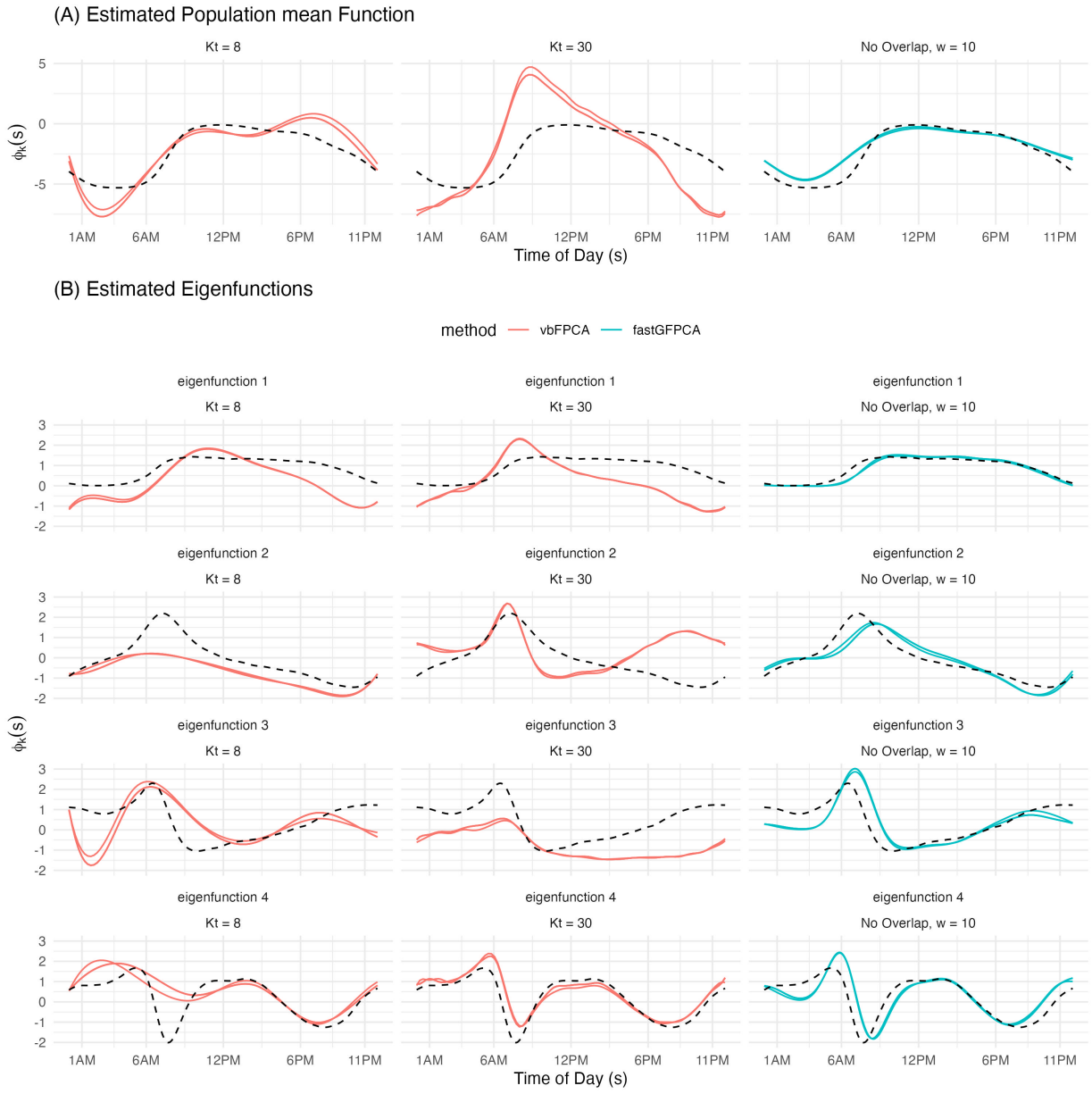


图 S3：根据 NHANES fast-GFPCA 结果，展示了 5 个模拟数据集的估计的总体均值函数 ($\hat{}\mu(s)$) 和前四个估计的 $\hat{}$ 特征函数 ($\phi_k(s)$)。模型估计值以红线 (vbFPCA) 或蓝线 (fast-GFPCA) 表示。展示了使用 Kt=8 (左列) 或 Kt=30 (中列) 基函数的模型的 vbFPCA 估计值，并将其与使用大小为 $w_l = 10$ 的非重叠窗口进行估计的 fast-GFPCA 结果进行比较。

S-4.2 非周期函数数据的模拟结果 ta

虽然本文主要关注基于具有周期端点的数据所得到的模拟结果，但我们也实施了采用非周期曲线的模拟情景。这些情景的结果汇总如下表 S1 所示。

非周期曲线模型参数的积分平方误差

估计值	N	J	fast-GFPCA 非重叠分箱 w = 6	fast-GFPCA w = 10	fast-GFPCA w = 50	vbFPCA K = 8	vbFPCA K = 30	tsGFPCA K = 30
$\psi_(s)$	100	100	0.95	0.98	-	0.54	1.28	**0.53**
		500	0.55	0.49	1.41	**0.31**	0.52	-
		2000	0.63	0.5	**0.4**	0.52	6.85	-
	500	100	0.44	0.48	-	**0.16**	0.28	0.22
		500	0.24	0.18	1.08	**0.09**	0.14	-
		2000	0.21	**0.14**	**0.14**	0.36	6.08	-
	1000	100	0.34	0.42	-	**0.1**	0.17	-
		500	0.14	0.12	0.97	**0.05**	0.09	-
		2000	0.1	**0.06**	0.08	0.35	5.92	-
$\beta_0(s)$	100	100	11.82	13.57	-	16.29	28.51	**10.09**
		500	**12.94**	13.07	12.55	13.16	15.71	-
		2000	11.88	**10.74**	13.2	12.84	13.08	-
	500	100	1.68	**1.66**	-	3.71	6	2.24
		500	1.69	1.63	**1.59**	3.29	3.87	-
		2000	1.96	1.93	**1.87**	3.33	3.22	-
	1000	100	**0.73**	0.78	-	1.89	3.09	-
		500	0.77	0.76	**0.71**	1.5	1.77	-
		2000	0.83	0.74	**0.7**	2.21	1.76	-
$\eta_i(s)$	100	100	**2.31**	2.32	-	2.44	3.14	2.4
		500	**0.56**	**0.56**	0.58	0.58	0.73	-
		2000	0.16	0.16	0.17	**0.15**	0.19	-
	500	100	2.29	**2.3**	-	**2.3**	2.42	2.32
		500	0.59	0.59	0.61	**0.55**	0.58	-
		2000	0.16	0.16	0.17	**0.15**	**0.15**	-
	1000	100	2.3	2.31	-	**2.29**	2.34	-
		500	0.57	0.58	0.6	**0.55**	0.56	-
		2000	**0.15**	**0.15**	0.16	**0.15**	**0.15**	-

表 S1：各方法和模拟情景下 $\psi_(s)$ 和 $\hat{}\eta_i(s)$ 的平均积分平方误差（MISE），以及 $\beta_0(s)$ 的积分平方误差（ISE）。每行中，在相应模拟情景下 MISE 最低的方法以 **粗体** 显示。本表汇总的所有数据均在总体均值 $\beta_0(s) = 0$ 和非周期特征函数条件下模拟生成。"-"表示该模型未针对相应模拟情景进行评估。